Lecture Alfred Fessard

Neurosciences et IA : Que peut apprendre l'IA de la manière dont le cerveau calcule ?

SIMON J. THORPE

Centre de Recherche Cerveau et Cognition (CerCo), CNRS-Univ Toulouse

La Lecture Alfred Fessard 2024 a été donnée le vendredi 24 mai 2024 au Centre Paul Broca Nouvelle-Aquitaine de Bordeaux.

Simon Thorpe est un neuroscientifique franco-britannique qui a consacré la majeure partie de sa carrière en France, où il s'est installé après un parcours académique commencé au Royaume-Uni. Il a étudié la psychologie et la physiologie à Oxford, avant d'entamer son doctorat sous la direction d'Edmund Rolls, puis a effectué un postdoctorat avec Max Cynader au Canada avant de rejoindre le CNRS en 1983. Il s'est installé à Toulouse en 1993, où il a cofondé le laboratoire CerCo (Cerveau et Cognition), contribuant à la renommée internationale de la ville dans le domaine des neurosciences cognitives. Il a dirigé le CerCo de 2014 à 2021.

Fasciné par la rapidité du traitement de l'information dans le cerveau, il raconte avoir découvert durant sa thèse des neurones du cortex orbitofrontal du singe qui répondaient sélectivement à une nourriture favorite en moins d'un dixième de seconde. Des expériences en électrophysiologie chez le singe éveillé et en EEG (électro-encéphalographie) chez l'homme confirment un traitement très rapide de l'information visuelle. C'est le thème de son article le plus cité (Speed of Processing in the Human Visual System) paru dans Nature en 1996. Partant de ce constat, le raisonnement de Simon Thorpe est à la fois simple et implacable : l'information visuelle doit traverser une dizaine d'étapes de traitement en un dixième de seconde, il n'y a donc qu'une poignée de millisecondes à chaque étape pour intégrer le signal et produire une réponse. Selon Simon, cette contrainte temporelle proscrit les boucles feedback et le codage de l'information par le taux de décharge neuronal—trop lents. Il propose un système feedforward, avec un seul potentiel d'action par neurone, où l'ordre de décharge parmi les neurones représente le code de l'information.

Simon Thorpe se démarque par son esprit inventif et ses idées originales. Une autre de ces idées lui a valu un prestigieux financement ERC Advanced en 2012 : il postule qu'une large proportion de notre cortex est une sorte de « matière noire » corticale, des neurones silencieux qui ne s'activent que très rarement pour reconnaître un souvenir ancien, comme le générique d'un feuilleton qu'on regardait étant petit, et qu'on n'a

jamais revu depuis. Au-delà des neurosciences, ses idées s'étendent à des domaines comme l'économie, proposant sur son blog des réformes économiques ou des solutions aux problèmes climatiques.

Simon Thorpe a toujours mis son énergie et sa créativité au service de la communauté. Directeur du CerCo, de l'Institut des Sciences du Cerveau Toulousain (TMBI : Toulouse Mind and Brain Institute), Président de l'ancienne CID 44 du Comité National (Cognition, langage, traitement de l'information, systèmes naturels et artificiels), à chaque fois Simon a insufflé son désir de changer les choses pour le meilleur. Son parcours et ses idées ont été mis à l'honneur lors de la Lecture Alfred Fessard 2024, organisée par la Société des Neurosciences.

Rufin VanRullen
Centre de Recherche Cerveau et Cognition (CerCo), CNRS – Univ Toulouse

rufin.vanrullen@cnrs.fr

****

L'intelligence artificielle s'est développée très rapidement au cours des dernières années, et les systèmes de pointe peuvent surpasser les humains dans un large éventail de tâches difficiles. Il s'agit notamment de tâches visuelles telles que la reconnaissance et la catégorisation d'images et de tâches audio telles que le traitement de la parole.

Les systèmes d'IA peuvent également générer des images, des vidéos, de la parole et de la musique, et contrôler des systèmes robotiques complexes. Ces progrès vont avoir de profondes répercussions sur la société en permettant de remplacer les humains par des systèmes artificiels plus rapides, moins chers et plus fiables.

Dans cette conférence, je note que nombre de ces avancées peuvent être considérées comme directement inspirées de la manière dont le cerveau calcule. De nombreux systèmes d'IA sont effectivement de très grands réseaux de neurones artificiels simples. C'est ce qui ressort clairement de l'article historique publié en 2012 par Geoffrey Hinton et ses deux étudiants de l'université de Toronto, Alex Krizhevsky et Ilya Suskever, intitulé « ImageNet Classification with Deep Convolutional Neural Networks » (Catégorisation d'ImageNet à l'aide de réseaux neuronaux convolutionnels profonds), qui a battu les systèmes de vision artificielle conventionnels avec un simple réseau neuronal feedforward à 7 couches. Cet article, déjà cité plus de 160 000 fois selon Google Scholar, a déclenché la révolution de ce que l'on appelle "Deep Learning" ou apprentissage profond. Tous ces systèmes reposent sur le principe selon lequel chaque neurone calcule une somme pondérée de toutes ses entrées et fait passer le résultat par une fonction de transfert avant d'envoyer sa valeur d'activation aux neurones de l'étape suivante. Chaque entrée a une valeur d'activation définie par un nombre à virgule flottante, généralement compris entre 0 et 1, et un poids synaptique, également défini par un nombre à virgule flottante.

On pourrait effectivement modéliser les 86 milliards de neurones du cerveau humain de cette manière, et ce serait des centaines de milliers de fois plus puissant que le modèle original de Hinton et de ses collègues, qui ne comportait que 650 000 neurones. Le problème est que le budget énergétique nécessaire à la simulation d'un tel système serait astronomiquement élevé. Si nous voulions mettre à jour l'état de chaque neurone toutes les millisecondes en supposant une moyenne de 7000 connexions par neurone, nous aurions besoin d'environ 600 pétaflops - un pétaflop étant égal à 10^{15} opérations en virgule flottante par seconde. Il n'existe actuellement qu'une poignée de superordinateurs sur la planète capables d'effectuer ces calculs, et ils ont tous des bilans énergétiques d'environ 20 mégawatts. En revanche, le vrai cerveau ne consomme que 20 watts. Il est donc environ 1 million de fois plus économe en énergie! Quel pourrait être le secret de l'efficacité énergétique phénoménale de nos cerveaux ?

Certains, dont le lauréat du prix Nobel Geoffrey Hinton, pensent que le secret réside dans l'utilisation d'opérations analogiques. Plutôt que d'effectuer de coûteuses multiplications en virgule flottante, on pourrait obtenir un résultat équivalent en codant les valeurs d'activation synaptique et les poids par des tensions et des résistances.

Le rôle clé des spikes

Cependant, j'ai passé 35 ans à étudier une caractéristique essentielle du cerveau, absente de presque tous les systèmes d'IA actuels. C'est le fait que les vrais neurones n'envoient pas de valeurs analogiques - ils envoient des spikes – les potentiels d'action. La plupart des ingénieurs qui conçoivent les puces d'IA, y compris Bill Dally, scientifique en chef chez Nvidia, savent parfaitement que les vrais neurones envoient des spikes. Cependant, ils considèrent que les neurones utilisent les spikes pour envoyer des valeurs d'activation par l'intermédiaire d'un code de taux de décharge (rate coding) qui est désespérément inefficace. Dally a effectivement déclaré que Nvidia n'utilise pas les spikes parce que l'envoi d'une valeur de 8 bits nécessiterait jusqu'à 255 spikes. Il vaut mieux envoyer un nombre de 8 bits comme un octet de données.

La croyance selon laquelle les neurones utilisent un codage fréquentiel est également partagée par de nombreux neurophysiologistes, qui décrivent souvent les propriétés des neurones sous la forme d'un histogramme temporel post-stimulus (PSTH) qui représente le taux de décharge en fonction du temps, en faisant la moyenne du nombre de spikes sur de nombreuses présentations de stimulus.

Cependant, en 1990, j'ai proposé un moyen beaucoup plus efficace d'encoder l'information, qui ne nécessite pas plus d'un spike par neurone. En effet, la latence de réponse d'un neurone sensoriel dépend de l'intensité de la stimulation - des stimuli plus intenses entraînent des réponses à latence plus courte. Cela signifie qu'il est possible d'encoder des informations dans l'ordre de décharge d'un ensemble de neurones. Chose incroyable, ce fait était visible dans les tout premiers enregistrements de l'activité du nerf optique de l'anguille effectués par le neurophysiologiste Edgar Douglas Adrian, lauréat du prix Nobel à Cambridge, en 1927 (figure 1). Pourtant, ce fait a été presque totalement ignoré par les neurophysiologistes et beaucoup de modélisateurs pendant plusieurs décennies.

Plus tard, ce schéma de codage temporel a été la clé de la création de SpikeNet Technology, une start-up de haute technologie que j'ai créée en 1999 avec deux de mes étudiants de l'époque - Rufin VanRullen et Arnaud Delorme, qui sont devenus par la suite directeurs de recherche au CNRS. Au cours des deux premières années d'existence de SpikeNet Technology, nous avons découvert qu'il était possible de construire des systèmes logiciels extrêmement efficaces pour le traitement d'images en utilisant des réseaux de neurones à spikes avec plusieurs caractéristiques nouvelles que nous n'avions pas dévoilées à l'époque pour préserver la "sauce secrète" de la société. Tous les traitements ont été réalisés avec des neurones qui émettent soit un seul spike, soit pas de spike du tout. En outre, nous avons conçu une règle d'apprentissage en un seul essai ("one shot"). Cette règle nous a permis de créer des neurones extrêmement sélectifs à n'importe quelle zone de l'image, en créant simplement des connexions fixes à partir des neurones d'entrée les plus actifs. Ces connexions synaptiques sont unaires ("unary"), c'est-à-dire qu'elles existent et ont une valeur de un, ou qu'elles n'existent tout simplement pas. Cela

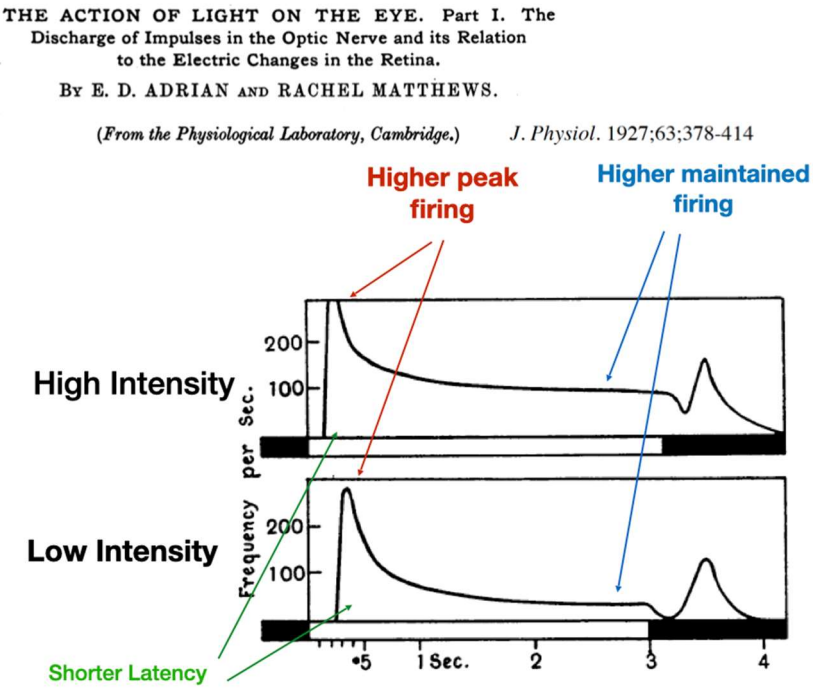

Figure 1. Adrian et Matthews (1927) ont comparé le taux de décharge des fibres du nerf optique de l'anguille pour un stimulus visuel de 3 secondes avec une stimulation de la rétine de forte ou de faible intensité. On peut voir que si les taux de décharge de pointe et de maintien sont plus élevés pour le stimulus plus lumineux, il y a également une réduction frappante de la latence de réponse.

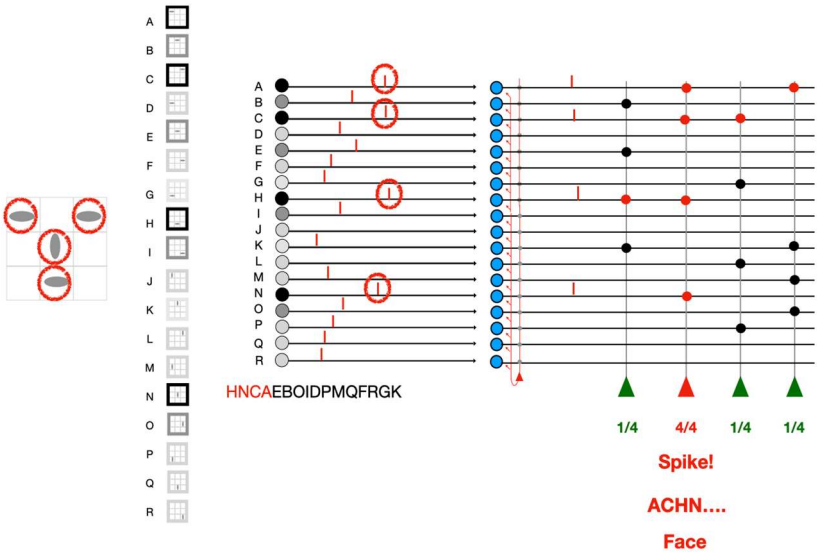

Figure 2. Un circuit simple de détection de visages avec 18 unités en entrée, correspondant aux bords horizontaux et verticaux d'une portion d'image de 3 x 3. Les neurones d'entrée utilisent un mécanisme de conversion intensité- latence tel que les neurones les plus actifs s'activent en premier. Ensuite, il y a des neurones relais contrôlés par un circuit inhibiteur qui intervient dès qu'un nombre donné de spikes a été reçu. Dans ce cas, ce mécanisme k-WTA (Winner-Take-All) ne laisse passer que les 4 premiers spikes. Les quatre neurones de sortie ont chacun 4 synapses avec un poids synaptique fixé à 1. Ici, le neurone connecté aux entrées A, C, H et N reçoit 4 “hits”, atteint le seuil et déclenche un spike signalant la présence d’un visage. Pour classer toutes les différentes façons de choisir 4 des 18 entrées, il faudrait 3060 neurones, chacun sélectif à un motif différent.

diffère des connexions binaires, beaucoup plus courantes, qui peuvent avoir une valeur de 0 ou de 1. Le mécanisme de base utilisé dans le moteur de reconnaissance de SpikeNet Technology est illustré dans la figure 2, qui montre un système simple avec 18 unités sélectives à l'orientation, codant les bords horizontaux et verticaux dans une simple portion de rétine de 3 x 3. Il convient de noter que, dans le système commercialisé par SpikeNet Technology, la zone d'image était plus grande (30 x 30 pixels) et qu'il y avait 8 orientations différentes au lieu de deux seulement.

Je pense que nous étions probablement une bonne décennie trop tôt, et pour diverses raisons, la société n'a pas vraiment réussi et a finalement été rachetée par BrainChip Inc en 2016. Brainchip a ensuite développé une puce basse consommation appelée Akida pour l'edge computing, mais n'a pas réussi à tirer pleinement parti de toutes les innovations que nous avions conçues au début des années 2000.

Vers des systèmes « Terabrain »

Cependant, plus récemment, j'ai revisité ces vieilles idées en développant une architecture que j'appelle Terabrain. La figure 3 illustre les principales différences entre le traitement de type Terabrain et les systèmes d'IA conventionnels qui effectuent des calculs en virgule flottante extrêmement coûteux.

Nous l’utilisons pour simuler de très grands réseaux de neurones à spikes qui peuvent être utilisés avec du matériel disponible sur le marché. Plus précisément, nous avons pu implémenter des réseaux comptant des milliards de neurones et des trillions de connexions synaptiques fonctionnant sur un simple MacBook avec un budget énergétique de quelques dizaines de watts – proche au cerveau humain ! La clé de cette efficacité remarquable réside dans le fait que les calculs ne sont nécessaires que lorsque les neurones émettent un spike. À partir de la liste des neurones qui se déclenchent à un moment donné, nous lisons la liste des neurones vers lesquels chaque neurone se projette à partir d'un grand fichier stocké sur un disque dur (SSD), puis nous ajoutons effectivement « un » au niveau d'activation de chaque neurone cible. Une fois que tous les spikes ont été propagés, nous sélectionnons les neurones ayant les niveaux d'activation les plus élevés et les faisons décharger lors du cycle de traitement suivant. Ce type de traitement est « sans zéro », en ce sens que nous n'utilisons jamais de zéros pour multiplier ou ajouter zéro à une valeur, ce qui nous permet d'éliminer une énorme quantité de calculs inutiles.

Il est important de noter que le nombre total de neurones dans le système n'a pratiquement aucune importance, car les seules limites réelles sont fixées par le nombre de neurones qui se déclenchent et le nombre de connexions que chaque neurone établit. Même avec un simple MacBook, nous pouvons avoir plus de neurones que le cerveau humain tout entier, car chaque neurone n'a besoin que d'un octet de mémoire pour mettre en œuvre une sorte de compteur simple. Un MacBook peut disposer de 128 Go de mémoire interne, ce qui permet d'implémenter 100 milliards de neurones de ce type – plus que le cerveau humain. Chacun de ces neurones peut avoir des centaines de connexions parce que les listes de connexions peuvent être stockées sur une mémoire SSD externe avec quelques octets pour chaque connexion. Dans de tels systèmes, les dispositifs de mémoire externes peuvent être reliés en chaîne à un seul port Thunderbolt, ce qui permet d'accéder à 80 To de mémoire, soit suffisamment pour

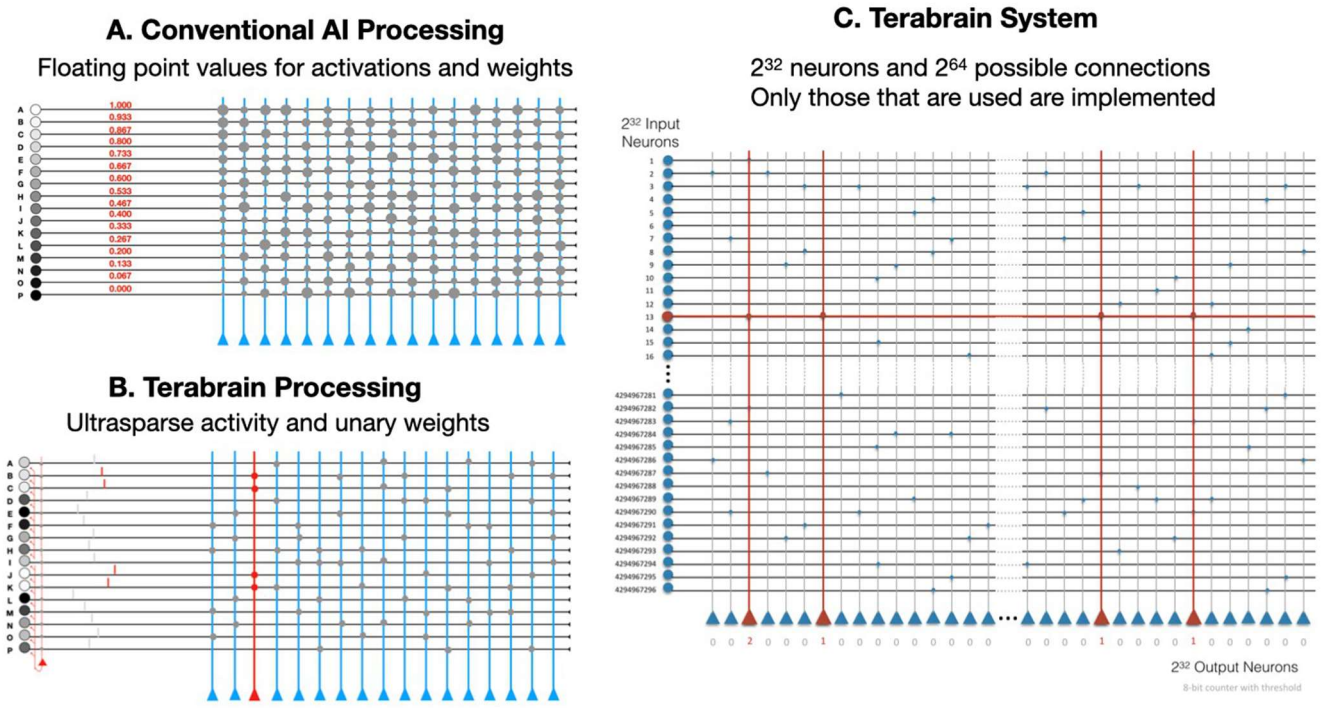

Figure 3. Le panneau A montre l'architecture neuronale utilisée dans l'IA conventionnelle. Les 16 unités d'entrée envoient chacune des valeurs en virgule flottante correspondant à leur niveau d'activation. Chaque neurone de sortie doit multiplier ces valeurs d'activation par une autre valeur en virgule flottante correspondant au poids synaptique. Le panneau B illustre le type de traitement utilisé dans Terabrain. Seul un petit pourcentage de neurones émet un spike à chaque cycle de traitement. Les poids synaptiques sont unaires. Le panneau C illustre un système Terabrain comportant plus de 4 milliards (2^{32}) de neurones. Cela équivaut à un “cross bar array” avec 2^{64} synapses. Mais comme seules les connexions existantes doivent être répertoriées, les besoins en matériel sont minimes. Une carte mémoire de 4 téraoctets peut stocker 1 000 milliards de connexions.

des milliards de connexions. Les dernières interfaces Thunderbolt 5 peuvent transférer des listes contenant des milliards de connexions par seconde.

Conclusion

Cela suffira-t-il à reproduire les performances du cerveau ? Un point essentiel est que, dans un tel système, la proportion de neurones de notre simulation qui émettent des spikes au cours d'une période d'une seconde peut être extrêmement faible. Cela peut sembler contradictoire avec le fait que la plupart des gens supposent que le taux de déclenchement au repos des neurones dans le cerveau est de 1 à 2 spikes par seconde. Cependant, je dirais que ce chiffre peut être complètement erroné car les neurophysiologistes ne peuvent fournir des chiffres que pour les neurones avec un niveau d’activité – les neurones silencieux sont invisibles. Plusieurs chercheurs ont souligné qu'un problème de « matière noire » peut se poser si une grande proportion de neurones n'a pas émis de spike pendant des heures, des jours, des semaines, voire des années. Il est probablement juste de conclure qu'à l'heure actuelle, la science n'a pas encore tranché. Mais il est possible que des techniques d'imagerie sophistiquées permettent de déterminer le véritable pourcentage de neurones silencieux. Si cette proportion est élevée, elle pourrait être un facteur important de l'incroyable efficacité énergétique du cerveau humain par rapport aux systèmes d'intelligence artificielle actuels. Cela ouvre la voie à la possibilité que l'activité neuronale ultra-éparse et les connexions unaires pourrait être la clé d'une deuxième révolution de l'IA, inspirée des astuces utilisées par les cerveaux biologiques.

simonjthorpe@mac.com