L'intelligence des réseaux de neurones

JEAN-PIERRE NADAL

Représentations simplifiées de parties du système nerveux des animaux, les réseaux de neurones résolvent des problèmes informatiques et éclairent certaines caractéristiques du fonctionnement cérébral.

Depuis 15 ans, des chercheurs de plusieurs disciplines – mathématiques, physique, biologie, etc. – étudient le comportement «intelligent» à l'aide de systèmes qui ont une architecture et un fonctionnement analogues à ceux de groupes de cellules dans le cerveau. Ces «réseaux de neurones formels», que nous désignerons désormais sous le nom de réseaux de neurones, résolvent des problèmes informatiques (reconnaissance de formes, par exemple) ou modélisent la cognition animale et humaine. Pour reproduire les mécanismes qui sont à l'origine de certaines capacités telles qu'apprendre, s'orienter, se déplacer, etc., on étudie le fonctionnement de réseaux de neurones à qui l'on fait exécuter des algorithmes, ou «calculs», qui semblent être ceux des groupes de neurones réels du cerveau.

Pour désigner ces calculs qui n'ont rien d'explicitement numérique, on emploie le terme de computation, mot d'origine latine, fort justement repris par la langue anglaise. Les sciences expérimentales sur lesquelles s'appuient ces travaux théoriques sont les neurosciences, notamment les neurosciences computationnelles ; ces disciplines tentent de relier l'organisation et les propriétés physico-chimiques des systèmes nerveux à leur fonction cognitive. La psychologie expérimentale, notamment la psychophysique et la psychologie cognitive participent également à cette recherche.

Naturellement, les spécialistes des neurosciences computationnelles ne prétendent pas que l'intelligence s'explique exclusivement par les propriétés de réseaux de neurones. Toutefois, il est bien démontré expérimentalement que des réseaux, c'est-à-dire des assemblées de neurones plutôt que des neurones individuels, jouent un rôle fonctionnel dans le cerveau. Par exemple, les neurophysiologistes ont démontré que, chez le rat, un groupe de neurones particulier code l'orientation de la tête ; l'activité de chaque neurone de ce groupe dépend de l'angle de la tête, mais celui-ci n'est correctement décrit que par l'activité de tous les neurones du groupe.

La première grande étude théorique des réseaux de neurones fut effectuée par le neurologue américain W. McCulloch et par le mathématicien W. Pitts en 1943. McCulloch reconnut ultérieurement qu'ils avaient voulu considérer le cerveau comme une machine universelle, au sens défini par le mathématicien britannique Alan Turing en 1937, c'est-à-dire un système capable de faire, en principe, n'importe quel calcul. McCulloch et Pitts montraient qu'on peut réaliser une machine de Turing avec des réseaux de neurones. Les «neurones» qu'ils envisageaient n'étaient pas de véritables cellules nerveuses, mais des modèles simplifiés de ces cellules : des «neurones formels», qui sont des «automates à seuil» ; ces petites unités qui additionnent les stimulations qu'elles reçoivent, produisent la valeur 0 ou 1 selon que cette somme est inférieure ou supérieure à un seuil *(voir la figure 2)*.

Certains psychologues considérèrent alors que l'on pouvait comprendre l'intelligence d'un cerveau par analogie avec les capacités d'un ordinateur. Cette idée reste aujourd'hui à la base de certaines approches de l'intelligence, en psychologie, en philosophie et en intelligence artificielle, mais l'étude des réseaux de neurones s'est développée, d'abord dans les années 1960, puis dans les années 1980, avec un point de vue tout différent.

En effet, un ordinateur classique a pour principale propriété d'être généraliste : on peut le programmer pour lui faire effectuer n'importe quelle tâche. Au contraire, un cerveau ressemble plutôt à un assemblage de machines spécialisées (le cortex visuel assure la vision, le cortex olfactif l'olfaction, le cortex moteur la commande des muscles...). Ces structures ont la faculté d'évoluer, de s'adapter à l'environnement. Les spécialistes des neurosciences computationnelles cherchent quelles sont les propriétés cognitives à la portée d'un réseau de neurones où la structure et la fonction sont intimement mêlées, où la frontière entre programme et matériel s'estompe. La faculté d'apprentissage étant l'une des plus fascinantes propriétés des systèmes vivants, la perspective d'en obtenir une modélisation par les réseaux de neurones est un attrait majeur.

La fonction dans la structure

La mémoire d'un réseau de neurones réside dans les poids, ou efficacités synaptiques, qui caractérisent les connexions entre les neurones : si un signal est émis par un neurone *A*, il est transmis à un neurone *B* avec une amplitude *a*, où *a* est ce que l'on nomme le poids de la connexion du neurone *A* au neurone *B*.

Dans les applications informatiques (par exemple, lors d'une tâche de recon-

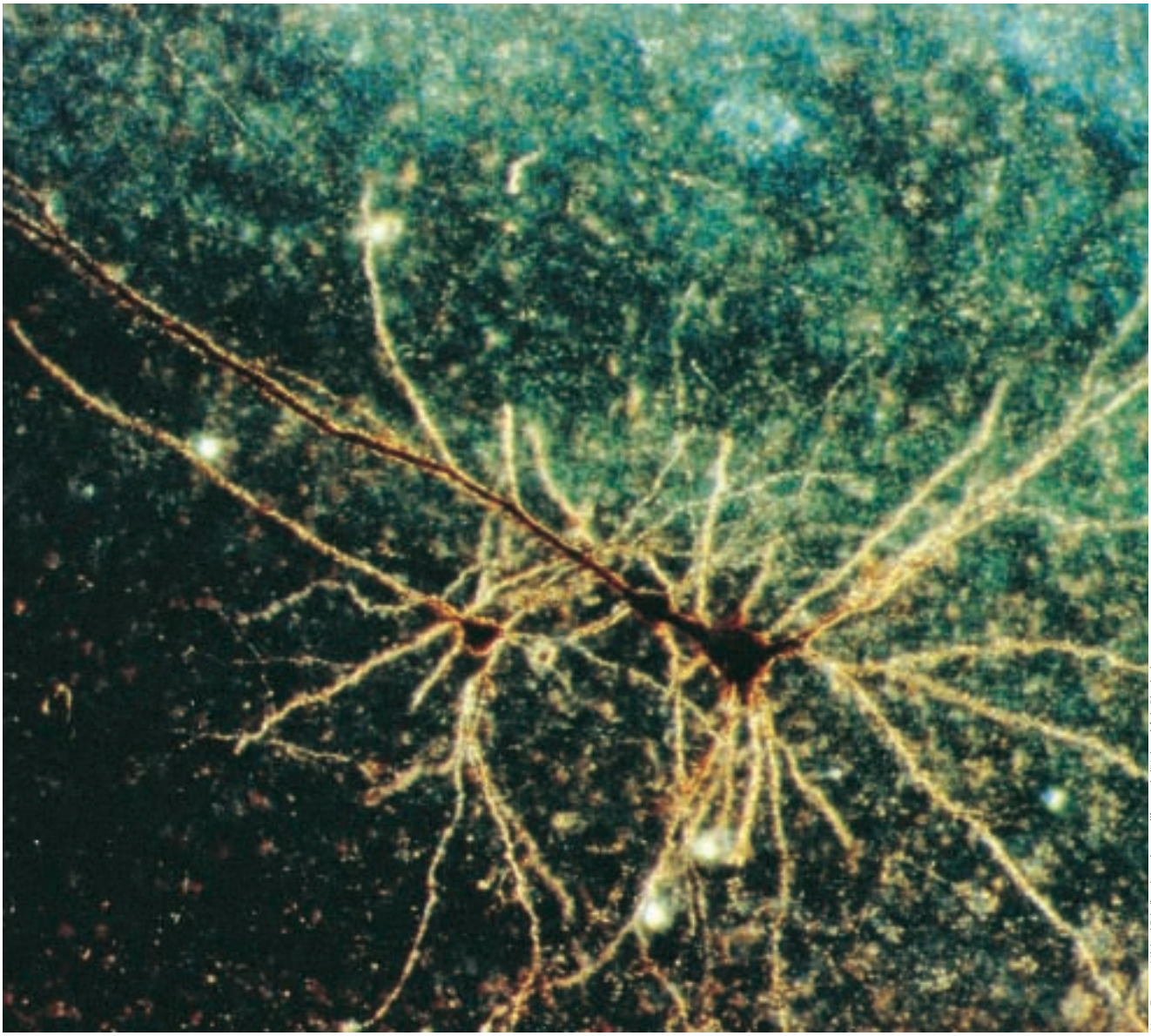
naissance de formes), l'apprentissage s'effectue par une modification de ces poids *(voir la figure 7)*. Initialement, on attribue des poids synaptiques aléatoires aux diverses connexions, et l'on applique aux neurones d'entrée un ensemble de valeurs. Le réseau calcule alors une réponse de sortie, qui diffère initialement de celle qu'il doit donner à la fin de l'apprentissage. On calcule un nouvel ensemble de poids synaptiques, afin que la sortie soit plus conforme à la sortie souhaitée. Progressivement, en modifiant ainsi les poids synaptiques alors que l'on présente au réseau un jeu de données, on obtient des poids qui permettent au réseau d'effectuer aussi bien que possible le calcul qu'on veut lui confier.

Cette procédure peut s'automatiser : on sait donner à un réseau la capacité de calculer lui-même les modifications des poids synaptiques qui s'imposent. Comme la fonction réalisée par le réseau, c'est-à-dire les «connaissances» et les compétences de ce réseau, est ainsi essentiellement déterminée par les connexions entre les neurones qui le constituent, on nomme parfois «connexionnisme» l'analyse des capacités cognitives des réseaux de neurones.

Nous considérerons ici deux aspects de l'«intelligence» des réseaux de neurones, tout deux concernant les facultés d'apprentissage des réseaux.

Réseaux à attracteurs et mémoire associative

La première propriété «intelligente» des réseaux de neurones concerne la mémoire. La mémoire humaine est différente de la mémoire d'un ordinateur classique, notamment en ce qu'elle est «associative». Elle nous permet, par exemple, de faire des mots croisés : une définition nous fait penser à un ou



Yves Fregnac, Michael Friedlander et leurs collègues de l'Université de Birmingham

1. LE CERVEAU est composé de neurones, qui se transmettent des signaux ; les signaux électriques qui se propagent à l'intérieur des neurones provoquent la libération de molécules, nommées neuromédiateurs, aux synapses. Les réseaux de neurones sont inspirés de cette architecture. Ici on voit deux cellules pyramidales du cortex visuel chez le cochon d'Inde ; la juxtaposition des prolongements cellulaires incite à penser que ces cellules sont connectées.

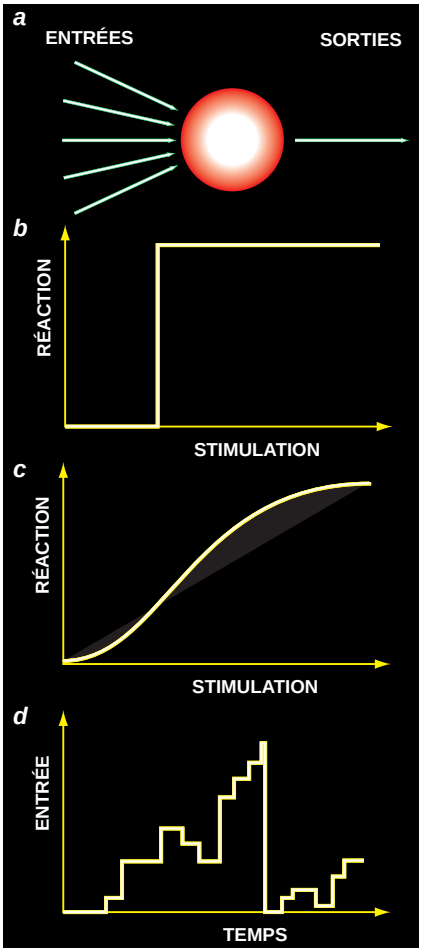
plusieurs mots, que nous essayons de placer dans la grille (une mémoire associative est également dite «adressable par le contenu»). Des physiciens ont découvert que les réseaux de neurones ont cette même propriété d'associativité.

En effet, les réseaux de neurones sont précisément des réseaux, c'est-à-dire des assemblées de neurones suffisamment interconnectés pour que l'activité de chaque neurone dépende des activités des autres neurones. Un cas extrême est celui d'un réseau à connectivité complète comme celui qui est représenté sur la figure 4 : quand on applique à chaque neurone du réseau une valeur d'entrée (c'est l'équivalent de la définition des mots croisés) et qu'on laisse ensuite le réseau évoluer, il se stabilise dans un état particulier, qui correspond au mot évoqué par la définition. Un tel réseau, s'il est suffisamment grand, peut ainsi avoir une fonction cognitive (par exemple, celle d'une mémoire associative) que ne peut posséder un neurone individuel.

La stabilisation d'un réseau ressemble à l'évolution de systèmes matériels : un matériau peut être dans différents états macroscopiques, ou «phases» (gazeux, liquide, solide), et les différentes phases ont des propriétés variées, telles que le ferromagnétisme (la capacité d'être aimantable, comme le fer). Ces états macroscopiques des matériaux résultent des interactions d'un très grand nombre de constituants élémentaires (atomes, molécules, moments magnétiques...).

Or, en 1982, à l'Université de technologie de Californie, le physicien John Hopfield a identifié une analogie fructueuse entre un groupe de neurones formels fortement interconnectés et un ensemble de moments magnétiques (ou «spins») en interaction. Cette analogie permet d'utiliser tout un savoir acquis en physique statistique ; elle permet de comprendre le rappel associatif, tel qu'il se produit dans la mémoire humaine, comme la convergence d'un système dynamique vers un «attracteur», et l'existence même de cette fonction de mémoire comme la manifestation d'un phénomène collectif. De même qu'un pendule simple finit par s'arrêter dans la position de repos, laquelle est un attracteur réduit à un point fixe, un réseau fortement connecté évolue jusqu'à un état particulier qui code le concept associé à l'information d'entrée (contrairement à un pendule simple, un réseau de neurones peut avoir un grand nombre d'attracteurs différents, chacun codant un objet mémorisé).

On retrouve avec ces modèles des caractéristiques décrites par la psychologie de la *Gestalt*, le courant développé au début du siècle par des psychologues allemands et de nouveau à l'honneur aujourd'hui : percevoir, c'est percevoir une forme *(Gestalt*, en allemand), un tout qui n'est pas la simple juxtaposition de ses parties (les éléments d'une scène visuelle, par exemple), mais qui résulte d'une interaction de ces parties. On peut dire, de même, que les attracteurs sont les seuls états d'activité du réseau qui portent un sens. Toute autre activité, y compris l'état initial directement imposé par le stimulus, n'apparaît que comme une étape intermédiaire dans la dynamique du réseau : une étape de calcul, qui mène à la réponse (l'attracteur associé).



2. DIVERS MODÈLES représentent les neurones de façon plus ou moins réaliste. Les neurones formels de McCulloch et Pitts produisent un signal de sortie égal à 0 ou à 1 selon les signaux d'entrée *(b)*. Les neurones sigmoïdes *(c)* ont une sortie continue, comprise entre 0 et 1, et d'autant plus proche de 1 que la somme pondérée des entrées est forte. Les neurones «qui intègrent et qui déchargent» *(d)* calculent la somme des stimulations et des inhibitions au cours du temps et déchargent quand cette somme devient supérieure à un seuil. Le modèle de M. Huxley et M. Hodgkin se fonde sur la physique de la membrane de la cellule nerveuse pour rendre compte des mécanismes d'intégration et d'émission de potentiels d'action.

Les interactions des neurones, c'est-à-dire les poids synaptiques qui quantifient la capacité des synapses à transmettre le signal, sont modifiées en fonction de l'activité du réseau, au cours de la phase d'apprentissage. Cet apprentissage conduit à la formation d'attracteurs qui représentent les objets à mémoriser. Qu'un état de mémoire soit un attracteur du système explique pourquoi une activité spécifique persiste en réaction à un stimulus, bien après que ce stimulus a disparu. Le fait que l'information codant un objet de mémoire soit répartie dans l'ensemble des connexions entre les neurones explique la grande robustesse des réseaux de neurones en cas de lésions.

Le modèle de J. Hopfield peut être considéré comme la formalisation mathématique la plus directe et la plus convaincante des idées avancées dans les années 1940 par Donald Hebb, de l'Université McGill. Ces idées restent à la base des travaux théoriques et expérimentaux sur la mémoire. Des études analytiques et numériques du modèle de J. Hopfield et de ses variantes ont conduit à une compréhension fine des mécanismes d'apprentissage dans de tels réseaux.

Du modèle à l'expérience

Ainsi, Daniel Amit et ses collègues physiciens de l'Université de Rome I, de l'Université hébraïque de Jérusalem et de l'École normale supérieure, à Paris, étudient depuis peu des modèles qui s'éloignent des systèmes magnétiques, pour mieux reproduire les interactions des neurones réels. Ils ont proposé de représenter les interactions de neurones par des potentiels d'action, c'est-à-dire des impulsions électriques brèves, plutôt que par des signaux continus. Comme dans le cerveau, ils considèrent que l'intérieur de chaque neurone présente une différence de potentiel par rapport au milieu extracellulaire

et que cette différence de potentiel est modifiée par les stimulations électriques reçues. Quand la différence de potentiel électrique dépasse un seuil, le neurone décharge : il émet une impulsion électrique qui se propage vers d'autres neurones, tandis que la différence de potentiel entre l'intérieur et l'extérieur du neurone revient à la valeur de repos.

Les modèles obtenus décrivent bien toute une série d'expériences de neurophysiologie nommées «tâches à réponse différée», effectuées par Y. Miyashita, de l'Université de Tokyo. Dans ce type d'expériences, un singe doit produire une action qui dépend du stimulus présenté un certain temps auparavant. Plus précisément, on montre à l'animal une image, engendrée par ordinateur, qui ne représente rien de familier ; puis, quand on lui montre une seconde image, on lui demande d'indiquer si cette image diffère de la première (le singe tient une manette ; il doit alors lâcher la manette et toucher un écran afin d'obtenir une récompense).

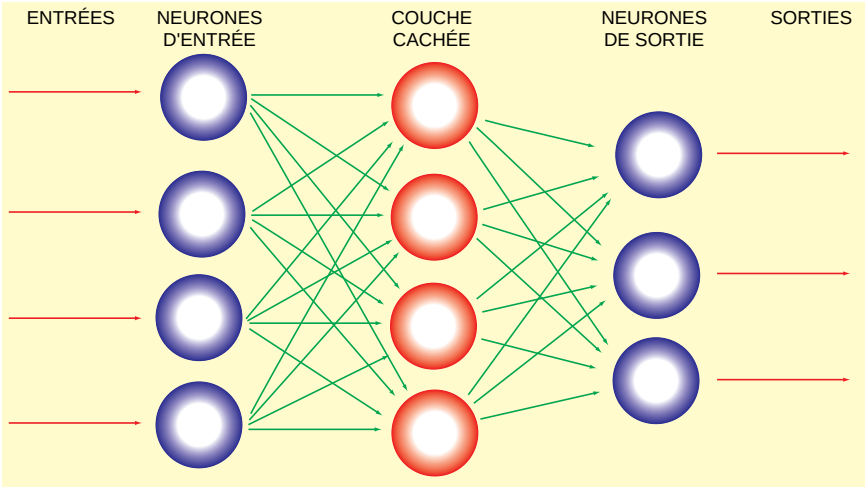
On enregistre, à l'aide d'électrodes implantées dans certaines aires corticales, tels le cortex préfrontal et le cortex inférotemporal, l'activité de neurones de ces aires. Les expériences montrent que, par endroits, les cellules ont une activité soutenue durant la période qui sépare le stimulus et la réaction. On constate que cette activité dépend sélectivement de un (ou plusieurs) des stimulus présentés durant l'expérience.

Ces données expérimentales sont en très bon accord avec les modèles de mémoire associative évoqués précédemment. En outre, les descriptions fondées sur ces modèles conduisent à des prédictions qui sont testées dans des expériences actuellement en cours à Jérusalem ; par exemple, quand les images sont présentées toujours dans le même ordre, on sait prédire certaines caractéristiques des configurations d'activités neuronales suscitées par les images. Pour la première fois, on cherche à montrer qu'un phénomène collectif est à l'origine d'une fonction cognitive.

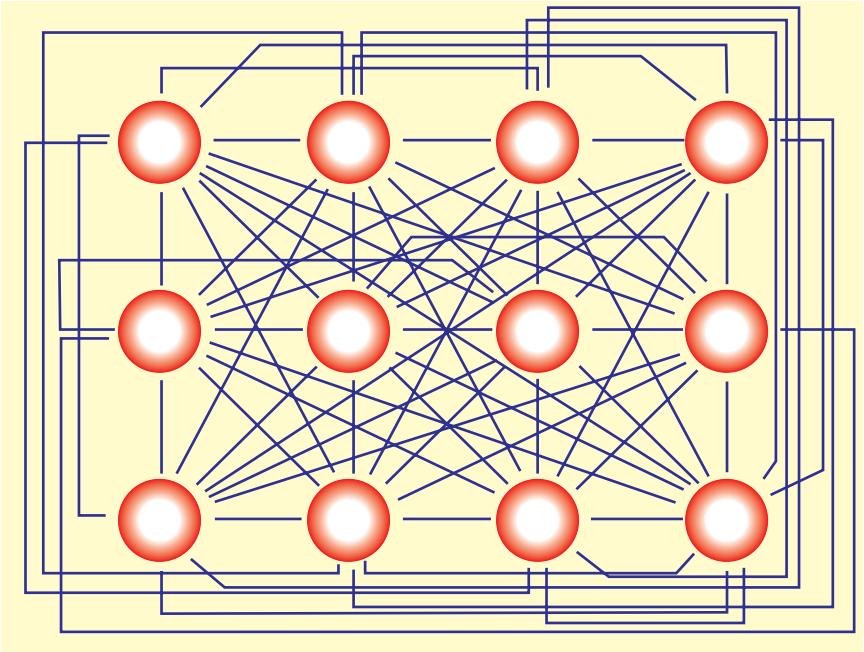
Apprendre et généraliser

La deuxième propriété «intelligente» qu'ont les réseaux de neurones est celle de «généralisation». Le cerveau humain généralise sans cesse, au point que les psychologues ont prétendu évaluer l'intelligence en mesurant cette capacité par les tests de quotient intellectuel (QI). Les tests classiques de QI comprennent beaucoup d'exercices où l'on demande de trouver la suite logique d'une séquence.

Comment poursuivre la série 1, 1, 2, 3, 5, 8, par exemple? Après un bref moment, on détecte la règle cachée : chaque nombre est égal à la somme des deux précédents, de sorte que le nombre suivant doit être 13 (la suite est nom-

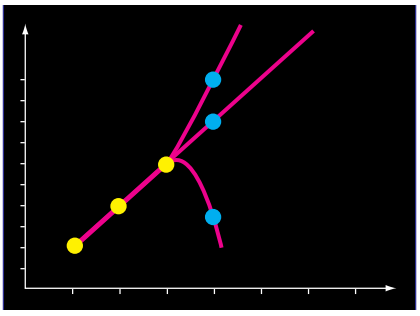


3. LES RÉSEAUX DE NEURONES EN COUCHES ont des propriétés qui dépendent du nombre de couches et du nombre de neurones par couche. Ce sont généralement ces types de réseaux que les informaticiens utilisent pour la reconnaissance de formes ; ils décrivent également de façon simplifiée les premières étapes de traitement des systèmes sensoriels (la rétine, par exemple).

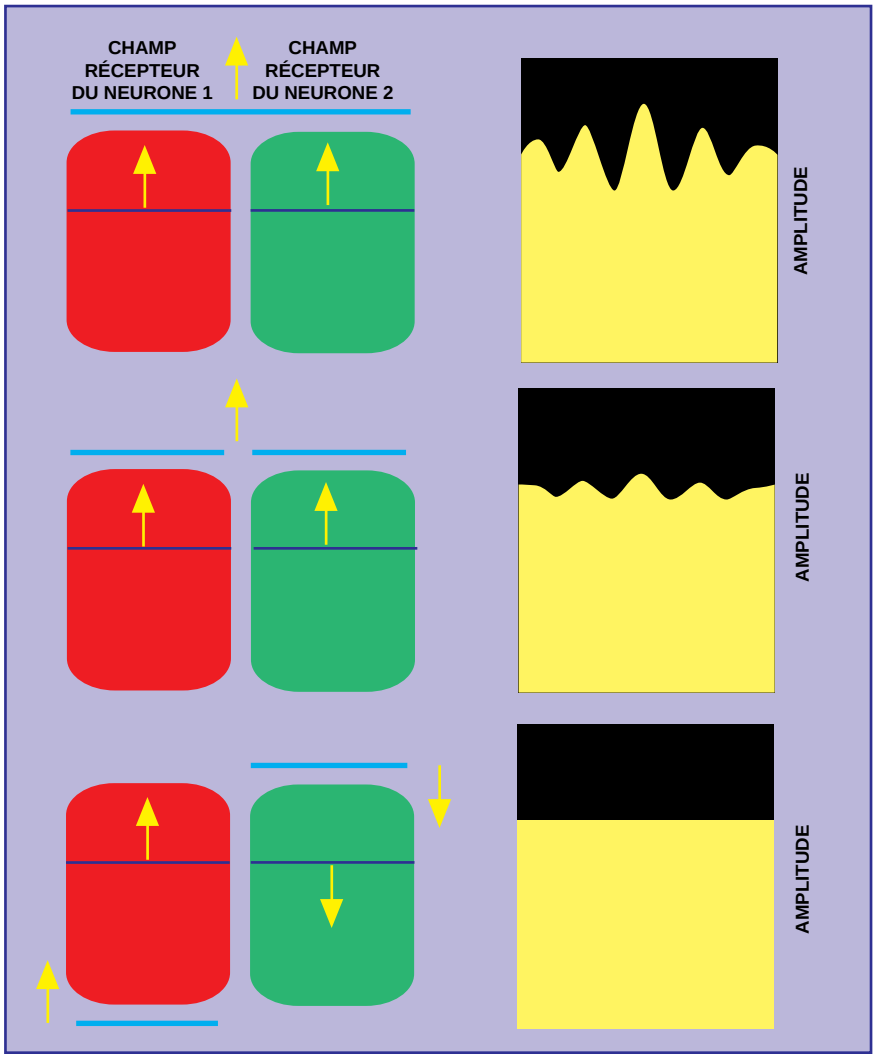


4. LES RÉSEAUX DE NEURONES COMPLÈTEMENT CONNECTÉS ont été étudiés dans les années 1980. Ils ont servi aux études de l'associativité de la mémoire, c'est-à-dire de la capacité analogue à celle qui nous permet de reconnaître une personne d'après son visage. Les théoriciens utilisent de tels réseaux avec un très grand nombre de neurones : on a déjà effectué des simulations avec 15 000 neurones, tandis que les parties de système nerveux à modéliser comportent quelque 100 000 neurones. Ces réseaux ont des capacités limitées, qui en font de bons candidats pour la description de la mémoire humaine à court terme. La nature de cette limitation est bien décrite par Sherlock Holmes : «Voyez-vous, le cerveau est comme un petit grenier d'abord vide. L'erreur est de s'imaginer que ce petit grenier a des murs extensibles. Soyez sûr qu'à un moment donné chaque nouvelle acquisition prend la place d'une ancienne. Il importe donc beaucoup de ne pas laisser les connaissances superflues évincer celles dont on a besoin.»

mée suite de Fibonacci). Toutefois on peut aussi bien prétendre que le nombre suivant est n'importe quel autre nombre, car une infinité de fonctions *f* prennent les valeurs 1, 1, 2, 3, 5, 8 pour les arguments 1, 2, 3, 4, 5, 6. La réponse *f*(7) = 13 nous paraît seulement plus «naturelle», plus «raisonnable» : on s'attend à ce que la règle cachée soit «simple». Peut-on comprendre cette faculté que nous avons à généraliser, autrement dit à inférer une règle à partir d'un nombre limité d'exemples? Peut-on comprendre pourquoi certaines généralisations paraissent plus naturelles, plus simples, que d'autres?



5. LA CAPACITÉ DE GÉNÉRALISER est utilisée dans les tests de QI. On présente une série de nombres telle que 2, 4, 6, et l'on demande le nombre suivant. Nous avons tendance à penser que ce nombre doit être 8, parce que les premiers termes sont les premiers nombres pairs de la suite. Pourtant, en théorie, n'importe quel nombre est possible, parce que plusieurs courbes peuvent passer par les points imposés. Par exemple, si l'on observe que 6 est égal à la somme des deux termes précédents, on peut ainsi poursuivre la suite par le nombre 10.



6. LA DYNAMIQUE des réseaux de neurones est une caractéristique importante de leur fonctionnement. Wolf Singer, à Francfort, a enregistré l'activité des neurones d'un singe à qui il montrait une barre *(en bleu clair)* qui se déplaçait. Il enregistrait alors des réactions synchrones de certains neurones (la courbe de corrélation à droite est accentuée quand les décharges des neurones sont synchrones) *(a)*. Deux barres mobiles excitant indépendamment les deux neurones suscitent une réaction inférieure *(b)*. Enfin, deux barres indépendantes qui se déplacent dans des directions opposées n'engendrent aucune réaction *(c)*.

L'analyse des réseaux de neurones a éclairé cette question. Mieux encore, ces réseaux de neurones ont été généralisés en «machines à apprendre» avec lesquelles on cherche à comprendre le fonctionnement du cerveau. De telles machines, réelles ou simulées sur ordinateur, sont décrites par quelques équations ; elles sont spécialisées dans la tentative de résolution de tâches analogues à la recherche de termes d'une suite. On admet qu'elles peuvent lire des questions, c'est-à-dire saisir des données dans un format prédéterminé ; puis elles répondent aux questions qu'on leur pose.

Par exemple, vers la fin des années 1980, une équipe des Laboratoires Bell de la Société *AT&T* a mis au point des réseaux de neurones pour la reconnaissance de chiffres manuscrits de codes postaux. L'équipe présentait des photographies numérisées de caractères manuscrits en entrée d'un réseau à trois couches cachées, et elle faisait apprendre au réseau à produire la sortie appropriée : le premier neurone de sortie devait être actif quand le réseau reconnaissait le chiffre *0*, le deuxième neurone devait être actif quand le chiffre *1* était reconnu, etc.

Plus de 7 300 chiffres étaient utilisés pour l'apprentissage. Après cette phase d'apprentissage, l'équipe présentait de nouveaux chiffres (environ 2 000) pour tester les performances en généralisation : le réseau afficherait-il la bonne sortie pour ces exemples non vus au cours de l'apprentissage? Le résultat fut assez satisfaisant : la proportion d'erreurs sur ces exemples tests était seulement de cinq pour cent (sur les mêmes données, un être humain fait moitié moins d'erreurs seulement). Plus récemment, avec un nombre de chiffres manuscrits bien supérieur (60 000 pour l'apprentissage et 10 000 pour le test), et avec des architectures et des algorithmes d'apprentissage plus complexes, la même équipe a obtenu un taux d'erreurs inférieur à un pour cent.

Dans ce type d'application, on modifie les poids synaptiques à l'aide d'algorithmes comparables à l'algorithme d'apprentissage précédemment évoqué. Toutefois, dans le cadre des neurosciences computationnelles, on considère

plus particulièrement des algorithmes d'apprentissage ayant une certaine plausibilité biologique : les poids synaptique sont changés selon un critère hebbien, c'est-à-dire en fonction des activités pré- et postsynaptiques.

Ainsi, après une phase d'apprentissage qui consiste à modifier les paramètres, une machine généralise – elle a compris la règle – si elle donne la réponse appropriée pour tout exemple qui n'a pas été considéré lors de l'apprentissage.

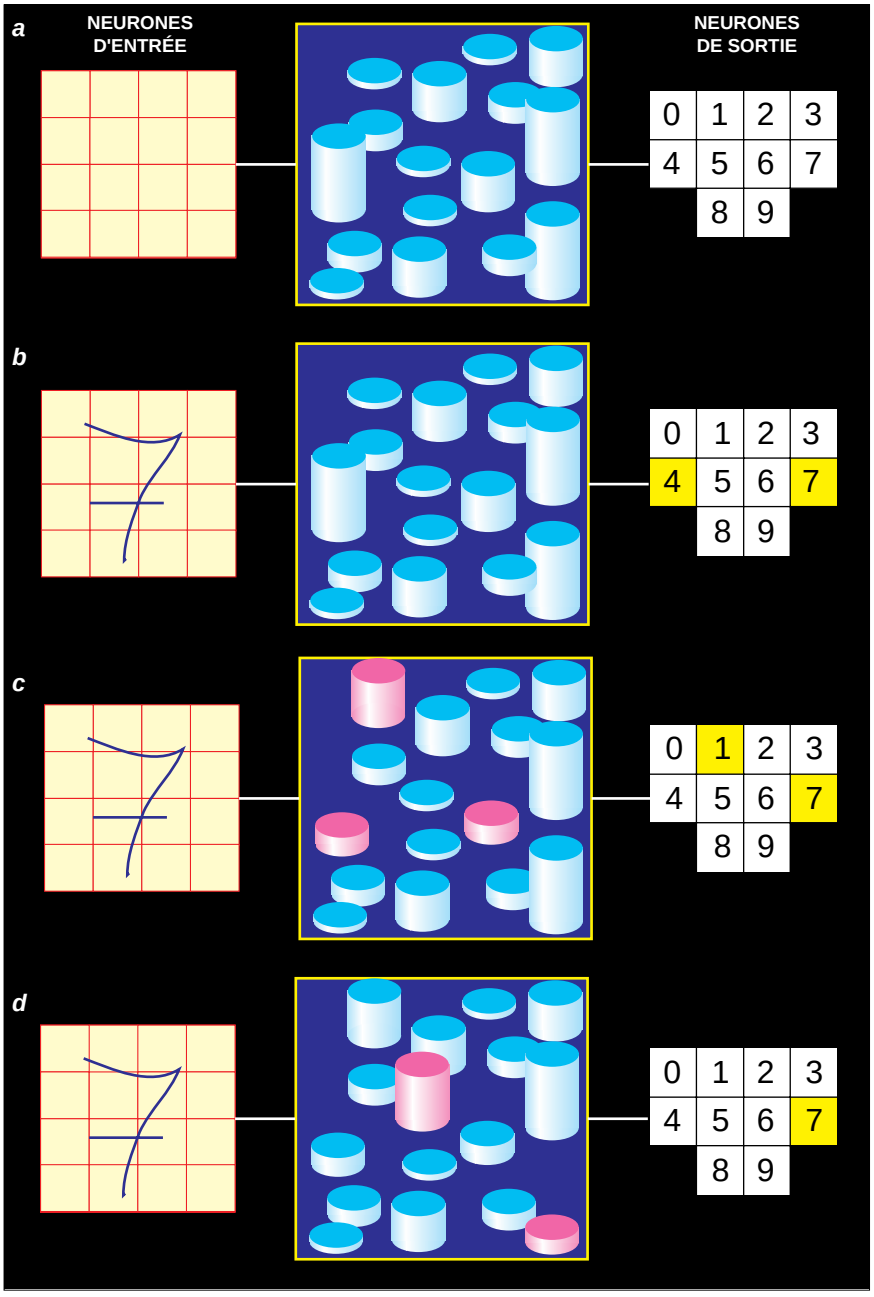
Questions de ressources

Supposons que l'on arrive effectivement à adapter les paramètres, de sorte que la machine produise la sortie désirée pour chaque exemple donné lors de l'apprentissage : ce serait la preuve que, pour chaque exemple, la machine a mémorisé l'association entrée-sortie désirée. Une telle mémorisation est possible si les ressources de la machine sont suffisantes, c'est-à-dire si l'architecture est appropriée et si le nombre de paramètres adaptables est suffisamment grand. Cependant, si ces ressources sont vraiment très grandes, la machine apprend par cœur tout nouvel exemple donné : quelle que soit la sortie désirée, il existera toujours un choix des paramètres qui permet d'obtenir cette sortie, tout en gardant les bonnes sorties pour tous les exemples déjà appris. La machine ne généralise alors pas, puisque à chaque instant elle peut mémoriser un nouvel exemple.

En revanche, si les ressources de la machine sont limitées, la mémorisation d'un grand nombre d'exemples risque d'encombrer la mémoire, de sorte que, lors de la présentation d'un nouvel exemple, on ne peut plus modifier les paramètres sans remettre en cause l'apprentissage déjà effectué. Autrement dit, avant que soit donnée la réponse à reproduire (la sortie désirée), la machine a déjà une réponse prête, la seule possible compte tenu de ce qu'elle a déjà appris. En ce sens, la machine généralise.

Plus on donne d'informations sur la règle à découvrir, plus on restreint le domaine des valeurs admissibles pour les paramètres, mais plus on limite le nombre de règles nouvelles que la machine peut encore apprendre. Si l'architecture est adaptée, la machine peut apprendre la règle cachée : il existe au moins un choix des paramètres qui permet à la machine de reproduire la règle ; dans ce cas, la généralisation sera correcte. Sinon, la machine généralise, mais infère une mauvaise règle : la règle trouvée ne coïncide que partiellement avec celle qu'il fallait trouver.

Ainsi la capacité à généraliser, à catégoriser, commence là où s'arrête la capacité à apprendre par cœur, et un système aux ressources illimitées ne peut qu'apprendre bêtement. Cette notion de ressources est subtile, car on



7. LA RÉTROPROPAGATION DU GRADIENT est un algorithme d'apprentissage très utilisé. Son principe est schématisé ici. Après le choix de l'architecture du réseau, par exemple un réseau en couches avec une couche cachée *(non représenté)*, on modifie les connexions entre les neurones *(au centre, on a représenté les poids synaptiques par des plots de différentes hauteurs)* pour que le réseau affiche une sortie désirée. On définit une fonction dite de coût, qui décrit la différence entre l'activité de sortie et l'activité désirée. Cette fonction est nulle si la sortie est celle que l'on désire. On présente alors des exemples au réseau, et, à chaque présentation, on modifie les connexions, de sorte que la sortie se rapproche de la sortie désirée. Par exemple, on présente ici un chiffre 7 ; le réseau réagit d'abord en affichant un 4 et un 7. On modifie alors les poids synaptiques *(en rouge, c)* et l'on présente à nouveau le chiffre 7 : le réseau réagit en affichant 7 et 1. Puis, quand on modifie encore les poids *(d)*, le réseau détecte le 7. De proche en proche, la sortie se rapproche de la sortie voulue.

sait réaliser des systèmes dont un seul paramètre est modifiable et qui, pourtant, ont des ressources infinies dans le sens précédent (c'est-à-dire une capacité infinie à mémoriser).

Considérons, par exemple, les nombres réels 0,8 ; 2,2 ; 5,1, et cherchons à les classer en deux classes. En considérant le signe du sinus de leur produit par un nombre *f*, on parvient à grouper ces nombres comme on le désire. Par exemple, si *f* est égal à $\pi/2$, 0,8 et 5,1 seront dans la première classe, tandis que 2,2 sera dans la seconde. Si *f* est égal à π, 0,8 et 2,2 seront dans la première classe et 5,1 dans la seconde. Ainsi, en choisissant convenablement *f* (le paramètre ajustable), on arrive à réaliser n'importe quel classement au choix. Ce procédé qui n'utilise qu'un seul paramètre s'applique généralement à un nombre quelconque de nombres à classer

Pour estimer les ressources, on utilise aujourd'hui un nombre de paramètres effectifs nommé dimension de Vapnik-Chervonenkis, du nom des deux chercheurs russes qui ont introduit cette notion dans les années 1970. Cette dimension est parfois infinie, comme dans l'exemple précédent, mais, dans la plupart des cas et, notamment, pour les réseaux de neurones, elle est de l'ordre de grandeur du nombre de paramètres modifiables. Notons que, pour la plupart des machines à apprendre, on ne sait pas calculer exactement cette dimension...

Une machine apprend par cœur tant que le nombre d'exemples à apprendre est inférieur à sa dimension de Vapnik-Chervonenkis, et elle commence à généraliser au-delà. Un apprentissage efficace par un réseau de neurones demande donc une architecture suffisamment complexe pour que la règle à apprendre soit effectivement réalisable par le réseau, mais une dimension de Vapnik-Chervonenkis pas trop grande, pour que l'apprentissage ne nécessite pas un trop grand nombre d'exemples : il s'agit de trouver le bon compromis, le réseau dont la complexité a le bon goût d'être bien adapté à la tâche à résoudre.

Ainsi, un réseau ou, plus généralement, une machine à apprendre qui a appris une règle est devenu un modèle pour cette règle ; il prédit le résultat de l'application de cette règle à tout nouvel exemple. On retrouve alors, dans le cadre de la théorie de l'apprentissage, la notion classique du rasoir d'Occam : le bon modèle (la bonne théorie) est celui dont la complexité est assez grande pour rendre compte des observations, mais pas plus.

Intelligence et sensations

Si nous avons privilégié ici les capacités d'apprentissage des réseaux de neurones, bien d'autres aspects de l'intelligence sont aujourd'hui étudiés à l'aide de réseaux de neurones. Ainsi de nombreuses équipes s'intéressent à la modélisation des systèmes sensoriels. D'autres modélisent les contrôles sensori-moteurs ou explorent des applications en contrôle de systèmes (pilotage d'un véhicule par un réseau de neurones, par exemple).

On étudie notamment les possibilités offertes par la dynamique des réseaux et par une activité neuronale qui est sous forme de potentiels d'actions, d'impulsions brèves. Ainsi des expériences en neurosciences indiquent que la mise en activité synchrone de groupes de neurones peut avoir un rôle fonctionnel important : par exemple, quand on présente une barre qui se déplace dans le champ visuel et que l'on enregistre l'activité de neurones du cortex visuel, on observe que des neurones éloignés ont pourtant une activité synchrone ; on a compris que ces neurones synchrones participent au codage de la barre dans le cerveau. Si l'on présente maintenant des morceaux de barre qui activent les mêmes neurones, on observe que ces neurones sont encore actifs, mais que leurs activités ne sont pas synchrones. Ainsi le synchronisme est une indication que les neurones réagissent au même objet *(voir la figure 6)*.

D'autres expériences de neuropsychologie, telles celles de S. Thorpe et de ses collègues de Toulouse, démontrent que la première vague d'impulsions engendrée par la présentation d'une image dans le champ visuel est suffisante pour détecter la présence ou non d'un animal dans une image. Dans certaines expériences, les neuropsychologues toulousains projettent à des sujets humains des images si rapidement que les sujets ne voient pas les détails, et ne peuvent qu'identifier globalement les images. On leur demande de maintenir un bouton enfoncé et de le relâcher quand un animal apparaît.

En mesurant le laps de temps entre la présentation des images et la réaction motrice des sujets, connaissant le temps de traitement dans chaque module cérébral ainsi que le temps de propagation des signaux nerveux, les psychologues ont montré que les réactions résultent d'un très petit nombre de traitements : on peut donc estimer le nombre de couches neuronales qui participent au comportement testé.

En mesurant l'activité du cerveau à l'aide d'électrodes posées sur le crâne (comme pour un électro-encéphalogramme), les neuropsychologues ont démontré que le système nerveux détecte la présence d'un animal dans l'image après seulement 150 millisecondes et que le cerveau n'utilise parfois que les toutes premières réactions des neurones stimulés.

Plus généralement, l'extrême efficacité algorithmique du système nerveux dans des tâches de reconnaissance de formes, qu'on ne sait absolument pas reproduire avec des modèles de réseaux de neurones classiques, pourrait résulter d'une utilisation astucieuse du codage par impulsions. Nous ne sommes pas à l'abri de très bonnes surprises à venir dans les toutes prochaines années...

Jean-Pierre NADAL est chercheur CNRS au Département de physique de l'École normale supérieure de Paris.

J. HERTZ, A. KROGH et R.G. PALMER, *Introduction to the Theory of Neural Computation*, Addison-Wesley, 1990.

G. HINTON, *Apprentissage et réseaux de neurones*, in *Pour La Science*, novembre 1992.

P. CHURCHLAND et T.J. SEJNOWSKI, *The Computational Brain*, MIT Press / Bradford Books, 1992.

J.-P. NADAL, *Réseaux de neurones : de la physique à la psychologie*, Armand Colin, 1993.

J. HÉRAULT et Ch. JUTTEN, *Réseaux neuronaux et traitement du signal*, Hermès, 1994.

M.A. ARBIB, *The Handbook of Brain Theory and Neural Networks*, Bradford Books / The MIT Press, 1995.

S. THORPE, F. FIZE et C. MARLOT, *Speed of Processing in the Human Visual System*, in *Nature*, vol. 381 pp. 520-522, 1996.