
APPRENDRE DANS LE TEMPS : UN NOUVEL ÉCLAIRAGE SUR L'APPRENTISSAGE CONNEXIONNISTE

Nicolas Szilas

LEIBNIZ-IMAG

INPG, 46, avenue Félix Viallet, 38031 Grenoble Cedex 01

Email : Nicolas.Szilas@imag.fr

Résumé

L'utilisation classique des réseaux connexionnistes, qu'il s'agisse de modélisation ou d'ingénierie, concerne l'apprentissage d'une et une seule tâche. Apprendre dans le temps, c'est au contraire considérer une succession temporelle de tâches auxquelles le réseau est confronté. Nous tentons de montrer l'importance de cette temporalité à grande échelle, aussi bien pour modéliser des comportements psychologiques que pour dépasser les limites computationnelles actuelles des réseaux connexionnistes. Pour cela, deux principes sont proposés : progressivité et polyvalence. Puis, leurs conséquences en terme d'optimalité sont discutées et amènent à postuler deux phases distinctes dans l'apprentissage, l'une étant l'acquisition et la coordination des habiletés, l'autre leur réorganisation.

1. Disciplines, méthode et objectifs

Les réseaux connexionnistes ont pour vocation d'une part de servir à modéliser les comportements humains ou animaux, et d'autre part de constituer de nouveaux algorithmes d'Intelligence Artificielle. Dans les deux cas, ils visent à constituer des systèmes cognitifs, ayant les propriétés suivantes :

- capacité d'apprentissage à partir d'exemples,
- représentation distribuée des connaissances,
- représentation implicite (i.e. souvent difficile à expliciter),
- traitements numériques.

En tant qu'alternative à l'approche symbolique, l'approche sub-symbolique développée par les réseaux connexionnistes constitue une voie prometteuse pour l'étude de la cognition, comme un certain nombre de chercheurs ont pu le mettre en évidence [Smolensky 88 ; Bates & Elman 92]. Cependant, au delà des caractéristiques générales de ces réseaux, il convient de constater un certain nombre de limites des algorithmes actuellement utilisés.

D'une part, sur le plan informatique, les réseaux actuels sont limités en efficacité : pour réaliser des apprentissages complexes, ils nécessitent un temps d'apprentissage croissant de façon exponentielle, qui devient vite un obstacle à toute simulation. Ce problème d'échelle ("scaling problem" en anglais [Haykin 94]) est dans la pratique allégué grâce à un choix très attentif des données par l'utilisateur du réseau, et grâce à une décomposition du problème en sous-problèmes et sous-réseaux.

D'autre part, lorsqu'ils sont utilisés en tant que modèles d'habiletés cognitives, ces réseaux demeurent extrêmement simplistes : ils ont peu d'entrées et sorties, les réseaux sont généralement homogènes (voire comportent deux modules [Jacobs *et al.* 91]), les types d'apprentissage sont souvent élémentaires (souvent de type behavioriste), les stimuli sont prétraités, etc. Une grande partie de la complexité des tâches réalisées par

les sujets humains n'est pas prise en compte dans les modélisations actuelles.

La position méthodologique de cette étude est de considérer que les deux limitations précédentes doivent se traiter de front. L'objectif est donc double : d'une part s'inspirer de données issues de la Psychologie pour améliorer les algorithmes existants et d'autre part disposer d'algorithmes connexionnistes plus intéressants, c'est-à-dire moins simplistes, pour modéliser les comportements humains. Ces deux objectifs correspondent aux deux limitations citées ci-dessus et sont étroitement liés : l'amélioration des algorithmes actuels passe par une prise en compte des contraintes psychologiques, tandis que l'étude de modèles psychologiques plus poussés ne peut se passer d'une réflexion de type informatique. Ainsi s'articulent autour de la présente étude deux disciplines, l'Intelligence Artificielle et la Psychologie Cognitive.

Les Neurosciences sont quelques peu écartées de notre recherche. En effet, d'une part nous soutenons la position selon laquelle il est possible de développer des modèles connexionnistes sans respecter strictement les contraintes neurobiologiques, car les connaissances actuelles concernant les mécanismes neuronaux ne permettent pas de modéliser les comportements de haut niveau qui intéressent psychologues et informaticiens, comme l'apprentissage de concepts ou l'automatisation. D'autre part, nous souhaiterions éviter le défaut classique qui consiste à "saupoudrer" l'argumentaire de "justifications" neurophysiologiques, sans aborder en profondeur le fonctionnement neuronal du système nerveux. Par prudence, nous préférons écarter les neurosciences de la discussion.

Pour atteindre les objectifs ci-dessus, le point central de la discussion va concerner le temps. Nous allons montrer que les réseaux opèrent sur un environnement relativement statique, négligeant la succession temporelle des situations d'apprentissage, et proposer de prendre comme point de départ cette temporalité à grande échelle. Cela nous amènera à poser deux principes fondateurs de l'apprentissage dans les

réseaux connexionnistes : progressivité et polyvalence. Ensuite, les conséquences de ces deux principes en terme d'optimalité seront analysées.

Précisons qu'il a été choisi de ne présenter dans cet article aucun algorithme informatique concret, ni même une architecture générale, car la réflexion qui suit est indépendante de tel ou tel réseau. Ici sont exposés des principes de conception, qui sont censés servir de base à de nouveaux algorithmes, alors qu'un algorithme est un exemple, une instantiation *parmi beaucoup d'autres* de ces principes. Ces principes méritent selon notre opinion une discussion à part entière.

2. Le temps des exemples dans les systèmes connexionnistes

Pour un réseau connexionniste, l'environnement est l'ensemble des exemples d'apprentissage qui lui sont présentés.

Dans une première utilisation d'un réseau supervisé (une sortie désirée est fournie au système), l'apprentissage se fait en temps différé sur un ensemble fini d'exemples présentés en un seul bloc au réseau, la plupart du temps à plusieurs reprises. Dans ce cas, il est clair que la dimension temporelle est inexistante : entre deux adaptations du réseau, le bloc d'exemple est le même, et l'intérieur d'un bloc, l'ordre des exemples n'a aucun effet.

Dans une seconde utilisation, l'adaptation du réseau se fait à chaque présentation d'un exemple, ou d'un petit groupe d'exemples. Chaque exemple, ou groupe d'exemples, est différent du suivant, mais ils sont tous extraits soit d'un même ensemble d'apprentissage fini (identique au bloc ci-dessus), soit d'un ensemble d'apprentissage infini, échantillonné aléatoirement selon une distribution fixe. Il apparaît que, même si l'exemple change à chaque itération, à une échelle plus grande, l'environnement est statique.

Il existe d'autres situations où les exemples sont dynamiques à une grande échelle : tâches de poursuites [Salganicoff 93], où le réseau doit modifier ses poids face à un environnement qui évolue ; apprentissage actif *informatif*, dans lequel un mécanisme sélectionne activement les exemples à apprendre, de manière à maximiser l'information issue de l'environnement [Atlas *et al.* 90 ; Hwang *et al.* 91 ; Krogh & Vedelsby 95]. On notera cependant que l'évolution de la distribution des exemples est lente.

Enfin, d'autres travaux, qui vont être exposés dans le détail à la section suivante, ont étudié l'effet de variations brusques de l'environnement, correspondant au fait que le réseau est impliqué dans une succession de tâches. Nous allons tenter de montrer que dans ce type de temporalité existe un facteur essentiel de l'apprentissage.

3. Apprendre petit à petit pour apprendre mieux

Ce paragraphe va montrer l'importance, sur le plan computationnel, de présenter les exemples regroupés dans des tâches successives distinctes, dans le sens d'un apprentissage à complexité croissante, que nous appelons par la suite *apprentissage progressif*.

Une des limitations des réseaux connexionnistes est le fait qu'ils tentent d'apprendre à partir du minimum de connaissances, à savoir une architecture la plus générale possible et des poids de connexions aléatoires [Geman *et al.* 92]. On note alors ce point de départ la *tabula rasa*. La première critique de cette *tabula rasa* est informatique : l'apprentissage est difficile, car le réseau doit apprendre un grand nombre de régularités uniquement dans les données. La seconde critique porte sur la plausibilité cognitive : est-il raisonnable de penser qu'un système cognitif tel que l'humain démarre sans aucune connaissance lorsqu'il apprend quelque chose ? L'utilisation de réseaux supervisés généraux, tels que les réseaux à couche à rétropropagation du gradient, selon la stratégie de la *tabula rasa* est donc aberrante, aussi bien au niveau informatique qu'au niveau psychologique.

La réponse classique à cette critique est la suivante :

Puisque la position méthodologique de la *tabula rasa* est fautive, il faut au contraire *partir de quelque chose*, en introduisant *a priori* des connaissances, sous forme de topologie du réseau [Bates & Elman 92], de compilation de règles [Towell *et al.* 91], de codage des entrées [Geman *et al.* 92], de connexions ciblées, de poids partagés [Le Cun 89]. En fait, toute application réaliste des réseaux connexionnistes procède de la sorte. Selon [Geman *et al.* 92], l'apprentissage en soi est un problème facile : la difficulté réside dans le choix des bonnes représentations a priori.

Mais introduire des connaissances a priori est finalement un retour en arrière, vers l'Intelligence Artificielle sans apprentissage, qui consiste à imposer au système artificiel une représentation a priori. Or on connaît les limites de cette approche : d'une part le système ne peut être conçu sans une forte connaissance du domaine auquel il s'applique ; d'autre part, la représentation imposée n'est pas nécessairement la meilleure pour le système.

Une autre justification de cette injection de connaissances a priori concerne la plausibilité cognitive. Sans entrer dans le débat inné/acquis, il est clair que l'approche de la *tabula rasa* correspond à une position complètement antinativiste peu acceptable : l'être humain naît sans aucune structure préformée. En injectant des connaissances a priori, on se retrouverait dans une situation plus fidèle à la réalité humaine [Bates & Elman 92]. Cette position n'est cependant pas acceptable, pour la plupart des utilisations des réseaux connexionnistes. En effet, si l'on apparie les connaissances a priori des réseaux et le ciblage cérébral avant la naissance, alors à une tâche d'apprentissage devrait correspondre soit l'apprentissage du nouveau-né, soit, pour des tâches adultes, l'apprentissage couvrant l'ensemble de la vie du sujet. Or, dans la plupart des cas, l'apprentissage est censé représenter une période du sujet de courte durée. Tout se passe comme si on oubliait toute l'histoire du sujet, entre sa naissance et le moment où l'apprentissage est mesuré.

C'est justement cette *histoire* du sujet, et du réseau, que l'on propose ici d'envisager.

Il est possible en effet de présenter à un réseau connexionniste un ensemble d'apprentissage jusqu'à ce qu'il l'apprenne, puis de présenter à ce même réseau un autre ensemble d'apprentissage pour qu'il l'apprenne à

nouveau, et ainsi de suite. L'idée est d'organiser cette succession d'ensembles d'apprentissage, correspondant à une succession de tâches, selon un ordre favorisant l'apprentissage, en allant du simple au difficile. On trouve aisément dans la vie de tous les jours des exemples de connaissances qui semblent ne pouvoir être apprises sans une décomposition en sous-tâches plus faciles : l'ensemble du cursus scolaire est fondé sur ce principe.

Des expériences menées dans plusieurs champs de la Psychologie ont confirmé quantitativement cette idée. En apprentissage moteur par exemple, plusieurs expériences décrites dans [Famose 90] montrent que si une tâche trop difficile est présentée au sujet, alors celui-ci n'arrive pas à l'apprendre, alors que proposer au sujet une succession de tâches dont la difficulté est graduée jusqu'à atteindre la même tâche complexe rend l'apprentissage possible.

En ergonomie logicielle, l'étude de [Caroll 84] montre que l'apprentissage d'un logiciel de traitement de texte est facilité si on décompose l'apprentissage en deux phases : dans la première, une version réduite de l'interface est présentée au sujet ; dans la deuxième, le logiciel complet est proposé.

Enfin, dans le domaine de l'apprentissage animal, pour faire apprendre à des animaux des comportements complexes, les psychologues ont souvent réussi à modifier le comportement, en augmentant la difficulté de la tâche très progressivement, par exemple la longueur de la séquence à mémoriser [Dorž & Mercier 92].

Sur le plan informatique, quelques expériences, rares, ont montré que l'apprentissage progressif était efficace dans les réseaux connexionnistes [Jacobs 88 ; Wieland & Leighton 88 ; Chen 90 ; Fahlman 91 ; Elman 93 ; Cloete & Ludik 93 ; Szilas & Ronco 95 ; Tetewsky *et al.* 95]. Deux de ces expériences sont plus convaincantes, dans le sens où l'apprentissage progressif permet non seulement d'accélérer l'apprentissage mais aussi et surtout d'augmenter très nettement la chance de réussite de l'apprentissage.

[Elman 93] expose une expérience de prédiction grammaticale utilisant un réseau récurrent, dans laquelle l'apprentissage échoue lorsque l'ensemble est présenté en entier. Par contre, il réussit si l'apprentissage est décomposé en sessions dont l'ensemble d'apprentissage correspondant contient au départ peu de phrases difficiles (enchaînements), mais de plus en plus dans les sessions suivantes.

Dans [Szilas & Ronco 95], un réseau unidirectionnel à couche entraîné par rétropropagation du gradient est confronté à la tâche de la double spirale [Fahlman & Lebiere 91], tâche de classification en deux dimensions, qui a la particularité et l'intérêt d'être difficile pour le type de réseau en question. Effectivement, sur 25 essais, seulement 4 se sont avérés réussis. En décomposant de manière spécifique l'ensemble d'apprentissage en 6 ensembles successivement inclus, l'apprentissage réussit 21 fois sur 25.

Il serait erroné d'affirmer que ces expériences contournent le problème de l'injection de connaissances a priori : l'ordonnement précis des tâches d'apprentissage est une forme d'introduction de connaissances a priori. Mais la différence fondamentale

réside dans le fait que ces connaissances sont introduites de l'extérieur, par l'environnement, et non par l'opérateur du système cognitif (le chercheur dans le cas des systèmes informatiques, l'évolution de l'espace dans le cas de l'homme).

Pour pouvoir dépasser l'obstacle de l'injection de connaissances a priori, il serait nécessaire de trouver un algorithme général capable de déterminer automatiquement le découpage de l'environnement en tâches successives de complexité croissante. Un tel algorithme fait encore défaut (un tel algorithme a été proposé dans [Cloete & Ludik 94] mais s'applique dans un cas particulier), et c'est certainement pour cette raison que l'apprentissage progressif est peu abordé.

Par ailleurs, on peut se demander si les réseaux constructifs (voir [Fiesler 94] pour un rapide état de l'art) ne réalisent pas implicitement un apprentissage progressif : en commentant l'apprentissage avec un réseau de petite taille, l'espace des solutions est réduit, et l'apprentissage serait plus facile ; il ne serait donc pas nécessaire de connaître a priori la structure du problème. Ainsi, la question suivante se pose : "la temporalité de la structure du réseau est-elle suffisante pour réaliser l'équivalent d'un apprentissage progressif ?".

Quelques expériences informatiques ont montré un effet positif de l'apprentissage progressif sur des réseaux constructifs [Chen 90 ; Fahlman 91 ; Tetewsky *et al.* 95], et tendent ainsi à prouver que la temporalité de la structure ne remplace pas celle des exemples, mais que les deux temporalités exhibent des phénomènes qui se combinent.

Sur le plan de l'apprentissage humain, on a montré que les réseaux constructifs reproduisent mieux certaines données de la Psychologie développementale que des réseaux statiques [Shultz *et al.* 94]. Est-ce à l'émulation qui permet à l'humain d'apprendre progressivement ? Il est actuellement difficile de répondre, car d'autres mécanismes pourraient "filtrer la complexité" de l'environnement, comme par exemple la capacité limitée de la mémoire de travail [Szilas & Ronco 95].

Nous laissons donc ce problème ouvert, tout en remettant l'hypothèse que pour un environnement complexe, le fait d'avoir un réseau simple ne suffit pas à filtrer la complexité.

Cette partie a mis en évidence le principe de *progressivité* sur lequel les réseaux connexionnistes devraient s'articuler, à savoir la succession temporelle des tâches d'apprentissage. La partie qui suit est consacrée au deuxième principe que nous souhaitons introduire.

4. Des réseaux polyvalents

On appelle "habileté" d'un réseau ("skill" en anglais) sa capacité à fournir une réponse appropriée pour un ensemble de stimuli correspondant à une situation donnée. Dans un cadre informatique, cette situation correspond à une *utilisation* d'un réseau, tandis qu'en modélisation cognitive, il s'agit plutôt d'un *contexte*. Nous distinguons ici la notion d'habileté de celle de tâche, qui correspond à l'apprentissage d'une habileté.

Nous dñommons ržseaux polyvalents les ržseaux dotžs de plusieurs habiletžs ^ un instant donnž.

Par exemple, dans un contexte informatique, un ržseau qui sait ^ la fois reconna"tre des caract• res et contr™ler un processus industriel est polyvalent. Sur le plan de la modžlisation, un ržseau polyvalent saurait reproduire plusieurs facultžs humaines comme par exemple lire et řcrire. Il s'av• re que la grande majoritž des mod• les connexionnistes, ou tout du moins les utilisations qui en sont faites, est dždiže ^ une seule t%che d'apprentissage, et poss• dent donc une et une seule habiletž, qu'il s'agisse de modžlisation psychologique [McClelland 89 ; Shultz *et al.* 94] ou d'Intelligence Artificielle.

Les expžriences relatžes à la section pržcždente ont portž sur des ržseaux successivement confrontžs ^ plusieurs t%ches durant l'apprentissage. Mais ces ržseaux sont dirigžs vers une habiletž finale unique, selon un unique canal de traitement ("single channel processing" [Clark & Thornton 96]). A la fin de l'apprentissage, les habiletžs intermždiaires sont soit effacžes, soit inutilisžes. Cette caractžristique est džnommže apprentissage indžpendant dans [Pratt 95], par opposition ^ l'apprentissage sžquentiel.

Seuls les travaux initižs par R. Caruana [Caruana 93] (voir aussi [Silver & Mercer 96]) montrent l'intžr• t computationnel des ržseaux polyvalents : un ržseau est en m• me temps entra"nž sur plusieurs t%ches, chacune correspondant ^ une unitž de sortie d'un traditionnel ržseau ^ couches. L'auteur montre expžriementalement qu'un ržseau confrontž ^ une t%che principale en m• me temps qu'un certain nombre de t%ches auxiliaires, a priori reližes ^ la t%che principale, apprend plus rapidement que dans la situation o• seule la t%che principale est pržsentže. Mais cette řtude se restreint au cas o• toutes les habiletžs sont acquises *en m• me temps*, jamais successivement.

Pourquoi poser la polyvalence comme deuxi• me principe fondateur de l'apprentissage dans les ržseaux connexionnistes ? Tout d'abord, de toute řvidence, l'homme est polyvalent et tire profit de cette polyvalence. Il poss• de de nombreuses habiletžs et chacune d'elles n'est pas acquise de mani• re indžpendante : l'ensemble des activitžs džj" acquises influent, positivement ou nžgativement, sur l'apprentissage d'une nouvelle habiletž. Ainsi, la problžmatique du *transfer* entre t%ches a řt largement řtudiže en Psychologie [Robins 96], qu'il s'agisse de mžmorisation, de ržsolution de probl• mes, de raisonnement par analogie [Gick & Holyoak 83], d'acquisition d'habiletžs motrices [Holding 87], etc.

Par ailleurs, sur le plan computationnel, les ržseaux polyvalents sont tr• s prometteurs, s'ils sont capables de ržaliser des transferts positifs entre t%ches. Il s'agit de penser un ržseau comme une accumulation d'habiletžs distinctes ; quand une nouvelle t%che d'apprentissage se pržsente, le syst• me a la possibilitž de ržutiliser les habiletžs passžes. Comparativement ^ l'apprentissage progressif, qui consiste ^ rechercher une sžquence d'habiletžs de difficultž croissante, l'approche polyvalente est beaucoup plus robuste : lors d'un apprentissage progressif, si la t%che pržcždente est mal choisie, l'apprentissage est bloquž, ^ cause de ce "canal unique" řvoquž plus haut, alors que pour un ržseau polyvalent, non pas une habiletž unique mais une palette d'habiletžs est disponible. Le džfi principal

des ržseaux polyvalents est de trouver des moyens et des crit• res pour džterminer quelles habiletžs ržutiliser et comment. Comme il va • tre discutž au paragraphe suivant, l'approche modulaire, dans laquelle le ržseau est en fait un ržseau de ržseaux est certainement la voie ^ suivre. On notera par ailleurs que cette idže de ržutiliser les connaissances passžes a džj" řt exploitže dans la branche de l'Intelligence Artificielle appelže "apprentissage symbolique" [Iba 89 ; Langley 95].

5. Discussion sur la notion d'optimalitž

Une des thžmatiques des ržseaux connexionnistes concerne le dilemme entre la plasticitž et la stabilitž : comment un syst• me peut • tre ^ la fois plastique, c'est-^-dire s'adapter aux nouvelles entržes, et stable, c'est-^-dire garder l'information accumulže [Murre *et al.* 92]. Par exemple, les ržseaux ^ couches classiques [Rumelhart & McClelland 86 ; Le cun 87], parce que leur structure est tr• s connectže, sont plastiques, mais peu stables : une nouvelle habiletž efface l'ancienne (probl• me de "l'oubli catastrophique" [French 91 ; Murre 92 ; Murre *et al.* 92]).

Apprendre dans le temps, selon le principe de progressivitž nous am• ne naturellement ^ ce dilemme entre la plasticitž et la stabilitž. Comme de plus, selon le principe de polyvalence, les habiletžs doivent • tre conservžes, alors le syst• me devra nžcessairement ne pas • tre trop plastique : une nouvelle entrže ne devra pas nžcessairement modifier tout le ržseau. Plus pržcisžment, chaque habiletž doit • tre codže dans une sous-partie du ržseau dont les poids doivent • tre plus ou moins gelžs apr• s apprentissage (forcžs ^ ne plus řvoluer). On nomme un tel ržseau un ržseau semi-distribuž, et on montre qu'il peut ržsoudre le probl• me de l'oubli [French 91]. Les ržseaux modulaires par exemple, qui peuvent spžcialiser un module ^ une habiletž, peuvent respecter les deux principes de progressivitž et polyvalence.

Un tel gel des poids, ou son corollaire la focalisation de l'apprentissage sur une sous-partie du ržseau, implique qu'il y a de fortes chances pour que la solution obtenue ne soit pas la plus simple possible. Par exemple, soit un ržseau confrontž ^ des t%ches boolžennes ; si ce ržseau apprend la fonction "ET", puis doit apprendre la fonction "NON ET", alors, comme les poids ont řt gelžs, la solution la plus rapide pour apprendre consistera ^ ržutiliser le sous-ržseau qui ržalise la fonction "ET", et inverser sa sortie par un traitement supplžmentaire. La structure obtenue comporte donc un double traitement, "ET" puis "NON", alors que l'apprentissage en partant d'une *tabula rasa* aurait donnž une structure plus simple, de m• me complexitž que le sous-ržseau qui ržalise la fonction "ET".

Ce constat remet-il en cause les principes de progressivitž et polyvalence discutžs ci-dessus ? La th• se de cet article est qu'au contraire, la recherche de la structure la plus simple, telle qu'elle sous-tend une certaine classe d'algorithmes, est ^ remettre en cause. Cela nous am• ne ^ analyser et discuter de la notion d'optimalitž.

L'approche classique de l'apprentissage connexionniste, issue des techniques statistiques de ržgression, consiste ^ rechercher le ržseau le plus

simple (ayant le minimum de degrés de liberté, c'est-à-dire le minimum de poids indépendants) pouvant apprendre l'ensemble d'apprentissage. On montre qu'un tel réseau est *optimal* du point de vue de la généralisation, c'est-à-dire la capacité à donner une réponse correcte pour un exemple non appris [Herz *et al.* 1991]. Intuitivement en effet, si on se place dans le cas extrême où le nombre de poids du réseau et le nombre d'exemples à apprendre sont du même ordre de grandeur, alors le réseau peut se contenter de mémoriser les exemples (apprentissage par cœur), ce qui compromet ses capacités à généraliser. Plus une représentation est compacte, plus elle constitue une abstraction des données réelles ; à l'extrême de la compacité, on trouve le symbole, capable de concentrer un grand nombre de situations dans une seule unité de représentation.

On peut néanmoins critiquer le propos ci-dessus en faisant les remarques suivantes :

- 1 — Premièrement, la redondance, qui augmente fortement la robustesse des réseaux connexionnistes, est exclue dans un réseau optimal : si on enlève une unité, les performances se dégradent sensiblement. Or la robustesse est justement un atout important des systèmes connexionnistes. Pour qu'un système soit à la fois optimal (au sens de minimal) et redondant, il faut le dupliquer, de manière que qu'une défaillance de l'un soit compensée par l'autre (l'image des systèmes électroniques ou informatiques qui ont besoin d'être reproduits à l'identique pour une utilisation en parallèle, dans les systèmes où la probabilité de panne doit être minimisée). Ce type de duplication est néanmoins à valider sur le plan psychologique.
- 2 — Deuxièmement, cette recherche de la structure minimale est généralement réalisée en maintenant le réseau le plus petit possible *tout au long de l'apprentissage*. Par exemple, la structure sera constante tout au long de l'apprentissage, ou bien, pour le cas des réseaux constructifs, sera très réduite au départ (uniquement les entrées et les sorties) pour ensuite augmenter sans jamais dépasser la structure minimale. Il convient aussi de rechercher d'autres techniques d'apprentissage où il est autorisé de dépasser la structure minimale, quitte à s'y ramener en fin d'apprentissage.

Ces deux remarques nous amènent à proposer le principe suivant : pour un problème complexe, l'obtention par apprentissage d'une structure optimale passe par une ou plusieurs phases où le réseau est non optimal, c'est-à-dire généralise mal.

Sur le plan computationnel, ce principe est en désaccord avec les approches classiques actuelles évoquées ci-dessus, où l'on pose comme objectif la recherche directe de la représentation optimale. En particulier, l'approche polyvalente mais non progressive, proposée dans [Caruana 93], est justement argumentée autour du fait que l'apprentissage simultané de plusieurs tâches permet de trouver une représentation plus générale, donc plus optimale. Notre argument est que la progressivité permet d'apprendre des tâches plus complexes, contrairement au réseau déterminé dans [Caruana 93].

Le principe de non-optimalité repose sur l'hypothèse que faire évoluer les paramètres du système connexionniste dans un espace plus grand, dans des zones où le réseau n'est pas optimal en terme de généralisation, facilite l'apprentissage. On aurait une sorte de "tape non optimale" nécessaire, avant d'aboutir à la structure optimale. Nous souhaiterions dans le futur pouvoir valider plus formellement cette hypothèse. La technique d'élagage, qui consiste à enlever des unités ou des connexions procédées de selon cette idée ; elle est souvent critiquée car elle oblige à manipuler un réseau trop complexe durant toute la phase d'apprentissage avant l'élagage.

Cette idée de passage par des structures non optimales est en accord avec certains résultats de la Psychologie cognitive et développementale : le développement d'habiletés cognitives chez l'enfant passe par des stades où celui-ci possède de une certaine connaissance pour résoudre un problème, connaissance que l'on peut qualifier de non optimale dans la mesure où elle n'est valide que pour certaines situations. Par exemple, au cours de l'apprentissage d'opérations mathématiques comme l'addition ou la soustraction, on a mis en évidence que les sujets commettent un certain nombre d'erreurs dites rationnelles, en ce sens qu'elles ne sont pas la résultante d'une réponse aléatoire, mais de l'application systématique d'une procédure que le sujet a acquise, procédure dont le domaine de validité s'avère limité [Ben-Zeev 95].

Poser que l'apprentissage doit passer par une phase non optimale a permis de maintenir les deux principes de progressivité et de polyvalence. Mais la question "comment obtient-on une représentation optimale ?" n'est en rien résolue, seulement repoussée. Il faut en effet envisager une deuxième phase d'apprentissage, qui permettra au système de passer d'une représentation non optimale à une représentation optimale. Quels types de mécanismes peuvent être alors mis en jeu ? Nous proposons deux idées directrices qui pourraient aider à trouver un tel mécanisme.

Premièrement, le passage vers une représentation optimale ne serait pas une transformation dans laquelle la structure nouvelle remplace et supprime l'ancienne. En effet, du point de vue du développement psychologique, on constate que d'anciennes procédures erronées peuvent resurgir alors que des procédures correctes sont déjà acquises. L'application de ces dernières se ferait en fait par inhibition des procédures erronées. Par ailleurs, il y a aussi un intérêt computationnel à garder les représentations sous optimales : d'une part cela permet d'utiliser une habileté pendant que celle-ci s'optimise ; d'autre part, la possibilité d'un système de pouvoir donner une réponse adaptée selon plusieurs mécanismes augmente sa fiabilité (on retrouve une idée voisine de la duplication mentionnée plus haut, mais il ne s'agit pas d'une duplication à l'identique).

Deuxièmement, le passage vers une représentation optimale implique certainement de profondes modifications du type de traitement effectué : il ne s'agirait pas d'altérations mineures de la représentation précédente. Prenons une tâche classique utilisée pour tester un réseau supervisé : la détection de parité. Cette tâche de classification binaire consiste à déterminer si le vecteur d'entrée possède de un nombre pair de valeurs à

1, à partir d'un ensemble d'exemples. Cette tâche se résout à l'aide d'un réseau à couches supervisées, mais le temps d'apprentissage augmente exponentiellement avec la dimension du vecteur d'entrée [Haykin 1994, p. 192]. L'étude de [Clark & Thornton 96] montre, dans le cas où le vecteur d'entrée est de dimension quatre, où l'ensemble d'apprentissage comporte seize exemples, qu'un réseau connexionniste n'a en fait aucune capacité de généralisation sur cette tâche : si seulement quinze exemples sont appris par le réseau, alors le réseau ne saura donner la bonne réponse pour le seizième. Ces deux études montrent que les réseaux utilisés trouvent une représentation pour résoudre le problème de parité qui n'est pas optimale : aucun de ces réseaux ne trouve une solution qui consiste à compter véritablement le nombre de valeurs à 1 dans le vecteur d'entrée. Une telle représentation est radicalement différente de la représentation sous optimale obtenue par le réseau (qui consiste à séparer l'espace d'entrée multidimensionnel par des hyper-plans).

L'exemple de la parité montre que ce n'est pas dans les données que se trouve la solution, mais dans une connaissance a priori que peut avoir le système, en l'occurrence l'action de compter. On retrouve la notion de transfert, discutée au paragraphe 4 à propos de la polyvalence, mais à un niveau supérieur, comme moteur de l'optimisation après apprentissage. C'est à ce niveau qu'un rapprochement avec la théorie psychologique de Karmiloff-Smith sur le développement cognitif pourrait selon nous se rattacher aux réseaux connexionnistes [Karmiloff-Smith 94]. Cette théorie expose le principe de la Redescription de la Représentation : un enfant, au cours de son développement, passe successivement d'une phase où il sait donner une réponse correcte dans une situation donnée à une phase où il *explícite* comment il donne cette réponse, ce qui lui permet de passer à un niveau supérieur de représentation. Cette Redescription de la Représentation correspondrait à la phase qui permet de passer d'une représentation non optimale à une représentation optimale, ou plus proche de la représentation optimale.

Il convient de noter que ce deuxième transfert (ou redescription) peut nécessiter l'attente d'un apprentissage ultérieur pour se produire. Pour reprendre l'exemple de la parité, l'action de compter peut être acquise après celle de la parité, et la représentation optimale ne pourra se former qu'après cette acquisition. Dans cette situation, même dans le cas extrême où l'apprentissage non optimal se réduit à un apprentissage par cœur, celui-ci est utile pour mémoriser la tâche en vue d'une restructuration ultérieure.

Conclusion

Nous venons d'ébaucher la voie qui permettrait d'élargir les potentialités actuelles des réseaux connexionnistes. Il s'agit de penser l'apprentissage dans le temps, et se pencher sur la dynamique des tâches, parallèlement à celle des représentations.

Deux principes ont été successivement introduits, progressivement et polyvalence. De ces deux principes découle le fait que le réseau devra être moins plastique que les réseaux homogènes, qui entraînent à son tour une non-optimalité du système. Cela a conduit à diviser l'apprentissage en deux phases distinctes : dans une première phase, les tâches d'apprentissage se succèdent et constituent un réseau d'habiletés, non

optimal ; dans une deuxième phase, ce réseau d'habiletés se restructure afin de se simplifier. Nous avons montré que ce raisonnement, qui peut s'effectuer uniquement sur le plan computationnel, correspond aussi à des réalités de l'apprentissage humain. C'est en cela que les algorithmes qui découleront de cette réflexion seront mieux adaptés à la modélisation psychologique. C'est aussi en cela que nous voyons dans ces principes, plus que des conseils pour la conception d'algorithmes, des principes généraux sur l'apprentissage.

Les perspectives de cette réflexion concernent en premier lieu l'implémentation d'un réseau connexionniste fondé sur les principes énoncés dans le présent article. Si aucun réseau existant ne convient, ne serait-ce que pour la première phase, c'est parce que l'architecture visée n'est en rien triviale : elle devra en effet combiner quatre caractéristiques, elles-mêmes loin d'être matrisées : temporalité de la structure [Fahlman & Lebiere 91], apprentissage multitâches [Caruana 93], transfert dans les réseaux [Sharkey & Sharkey 93 ; Pratt 95], apprentissage progressif [Szilas & Ronco 95].

Quant à la deuxième phase, l'obtention de mécanismes précis relève encore du défi. Il se peut que ces mécanismes fassent intervenir d'autres outils que le calcul neuronal, si l'on considère que la réorganisation qu'elle suppose nécessite une explicitation des traitements effectués, comme le suggère la théorie de Karmiloff-Smith.

Par ailleurs, quand il s'agira de confronter cette architecture informatique à des observations psychologiques, la situation sera plus délicate que dans le cas actuel des réseaux "mono-tâche" : on ne peut contrôler en laboratoire qu'un faible nombre de tâches. Un modèle polyvalent nécessitera de faire des hypothèses sur la nature de l'historique auquel un sujet humain a été confronté, avant telle ou telle expérience de laboratoire.

Références bibliographiques

- [Atlas et al., 1990] Lee Atlas, David Cohn, Richard Ladner, M. A. El-Sharkawi, R. J. Marks II, M. E. Aggoune & D. C. Park. Training Connectionist Networks with Queries and Sampling. In D. S. Touretzky (Ed.): *Advances in Neural Information Processing Systems 2*, pp. 566-573, San Mateo: Morgan-Kaufmann, 1990.
- [Bates & Elman, 1992] Elisabeth A. Bates & Jeffrey L. Elman. Connectionism and the study of change. Tech. report, CRL TR 9202, Feb. 1992.
- [Ben-Zeev, 1995] Talia Ben-Zeev, 1995. The nature and Origin of *rational Errors* in Arithmetic Thinking: Induction from Examples and Prior Knowledge. *Cognitive Science*, **19**, pp. 341-376, 1995.
- [Caroll, 1984] John M. Caroll. Minimalist design for active users. In B. Shackel (Ed.): *Human-computer interaction - INTERACT'84*, pp. 39-44, Amsterdam: North-holland, 1984.
- [Caruana, 1993] Richard A. Caruana. Multitask Connectionist Learning. *Proc. of the Connectionist Models Summer School*, pp. 372-379, 1993.
- [Chen, 1990] James R. Chen. A compositional Connectionist Architecture. In D. S. Touretzky (Ed.): *Advances in Neural Information Processing Systems 2*, pp. 188-197, San Mateo: Morgan-Kaufmann, 1990.

- [Clark & Thornton, 1996] Andy Clark & Chris Thornton. Trading Spaces: Computation, Representation and the Limits of Uninformed Learning. To appear in *Behavioral and Brain Sciences*.
- [Cloete & Ludik, 1993] Ian Cloete & Jacques Ludik. Increased Complexity Training. *Proc. IWANN'93*, Sitges, Spain, June 1993.
- [Cloete & Ludik, 1994] Ian Cloete & Jacques Ludik. Delta training strategies. *Proc. IEEE Conf. on Neural Networks*, pp. 295-298, Orlando, Florida, June 28 - July 2, 1994.
- [Dorž & Mercier, 1992] François Dorž & Pierre Mercier. *Les fondements de l'apprentissage et de la cognition*. P.U. Lille, Ga'tan Morin Editeur, 1992.
- [Elman, 1993] Jeffrey L. Elman. Learning and development: the importance of starting small. *Cognition*, **48**, pp. 71-99, 1993.
- [Fahlman & Lebiere, 1991] Scott E. Fahlman & Christian Lebiere. The Cascade-Correlation Learning Architecture. Tech. Report, CMU-CS-90-100, August 1991.
- [Fahlman, 1991] Scott E. Fahlman. The Recurrent Cascade-Correlation Architecture. Tech. Report, CMU-CS-91-100, May 1991.
- [Fiesler, 1994] Emile Fiesler. Comparative Bibliography of Ontogenic Neural Networks. *Proc. ICANN*, pp. 193-196, Sorrento, Italy, May 26-29 1994.
- [Famose, 1990] Jean-Pierre Famose. *Apprentissage moteur et difficultŽ de la t%oche*. INSEP, 1990.
- [French, 1991] Robert M. French. Using Semi-Distributed Representations to Overcome Catastrophic Forgetting in Connectionist Networks. CRCC Technical Report 51-1991, Indiana University, 1991.
- [Geman et al. , 1992] Stuart Geman, Elie Bienenstock & RenŽ Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, **4**, pp. 1-58, 1992.
- [Gick & Holyoak, 1983] M. Gick & K.J. Holyoak. Schema induction and analogical transfer. *Cognitive Psychology*, **15**, pp. 1-38, 1983.
- [Haykin, 1994] Simon Haykin. *Neural networks - A Comprehensive Foundation*. New-York: Macmillan College Publishing Company, 1994.
- [Hertz et al. , 1991] J.A. Hertz, R.G. Palmer & A.S. Krogh. *Introduction to the Theory of Neural Computation*. Redwood City (CA): Addison-Wesley, 1991.
- [Holding, 1987] D.H. Holding. Concept of training. In Gavriel Salvendy (Ed.): *Handbook of human factors*, pp. 939-962, New-York: John Wiley & Sons, 1987.
- [Hwang et al. , 1991] Jenq-Neng Hwang, Jai J. Choi, Seho Oh & Robert J. Marks II. Query-Based learning applied to Partially Trained Multilayer Perceptrons. *IEEE Trans. on Neural Networks*, **2**(1), pp. 131-136, January 1991.
- [Iba, 1989] Glenn A. Iba. A heuristic Approach to the Discovery of Macro-operators. *Machine Learning*, **3**, pp. 285-317, 1989.
- [Jacobs, 1988] Robert A. Jacobs. Initial Experiments On Constructing Domains of Expertise and Hierarchies In Connectionist Systems. *Proc. Connectionist Models Summer School*, pp. 144-153, 1988.
- [Jacobs et al. , 1991] Robert A. Jacobs, Michael I. Jordan & Andrew G. Barto. Task Decomposition Competition in a Modular Connectionist Architecture: The What and Where Vision Tasks. *Cognitive Science*, **15**, pp. 219-250, April 1991.
- [Karmiloff-Smith, 1994] Annette Karmiloff-Smith. PrŽcis of *Beyond modularity: A developmental perspective on cognitive science*. *Behavioral and Brain Sciences*, **17**, pp. 693-745, 1994.
- [Krogh & Vedelsby, 1995] A. Krogh & J. Vedelsby. Neural Networks Ensembles, Cross Validation, and Active Learning. In G. Tesauro et al. (Eds.): *Advances in Neural Information Processing Systems 7*, Cambridge (CA): MIT Press, 1995.
- [Langley, 1995] Pat Langley. Order Effects in Incremental Learning. In Reimann & Spada (Eds.) *Learning in Humans and Machines: Towards an Interdisciplinary Learning Science*, Elsevier Science, 1995.
- [Le Cun, 1987] Yann Le Cun. Mod• les connexionistes de l'apprentissage. Th• se de doctorat, UniversitŽ Paris VI, 1987.
- [Le Cun, 1989] Yann Le Cun. Generalization and Network Design Strategies. In R. Pfeifer, Z. Schreter, F. Fogelman-SoulitŽ & L. Steels (Eds) *Connectionism in perspective*, North Holland: Elsevier Science Publishers B.V, 1989.
- [McClelland, 1989] Parallel distributed processing: Implications for cognition and development. In R.G.M. Morris (Ed.) *Parallel distributed processing: Implications for psychology and neurobiology*, pp. 8-45, Oxford University Press, 1989.
- [Murre, 1992] Jacob M. J. Murre. The Effect of Pattern Presentation on Interference in Backpropagation Networks. *Proc. of the fourteenth Ann. Conf. of the Cognitive Science Soc.*, pp. 54-59, Chicago, Illinois, August 7-10, 1992.
- [Murre et al. , 1992] Jacobs M. J. Murre, R. Hans Phaf & Gezinus Wolters. CALM: Categorizing and Learning Module. *Neural Networks*, **5**(1), pp. 55-82, 1992.
- [Pratt, 1995] Lorien Y. Pratt. Using the discriminability based transfer algorithm to selectively bias neural network learning using the results of learning on related tasks. Submitted to *Journal of Artificial Intelligence Research*, 1995.
- [Robins, 1996] Anthony Robins. Transfer in Cognition. Submitted to *Connection Science*, 1996.
- [Rumelhart & McClelland, 1986] David E. Rumelhart & James L. McClelland (Eds). *Parallel Distributed Processing*, volume 1, Cambridge, Mass.: MIT Press, 1986.
- [Salganicoff, 1993] Marcos Salganicoff. Improved Learning of Time-Varying Mappings with Performance-Error Based Forgetting. Technical Report, Univ. of Pennsylvania, Dept. of Computer and Information Science, MS-CIS-93-80, Sept. 7 1993.
- [Sharkey & Sharkey, 1993] Noel E. Sharkey & Amanda J. C. Sharkey. Adaptive Generalization. *Artificial Intelligence Review*, **7**, pp. 313-328, 1993.
- [Shultz et al. , 1994] Thomas R. Shultz, Denis Mareshal & William C. Shmidt. Modeling Cognitive Development on Balance Scale Phenomena. *Machine Learning*, **16**, pp. 57-92, 1994.
- [Silver & Mercer, 1996] Daniel L. Silver & Robert E. Mercer. The Parallel Transfer of Tasks Knowledge Using Dynamic Learning Rates Based on a Measure of Relatedness. To appear in *Connection Science*, 1996.
- [Smolensky, 1988] Paul Smolensky. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, **11**(1), pp. 1-23, March 1988.
- [Szilas & Ronco, 1995] Nicolas Szilas & Eric Ronco. Action for learning in non-symbolic systems. *Proc. of the European Conference on Cognitive Science*, pp. 55-63, Saint-Malo, France, April 4-7 1995.
- [Tetewsky et al. , 1995] Sheldon Tetewsky, Thomas R. Shultz & Yoshio Takane. Training regimens and function compatibility: Implications for understanding the effects of knowledge on concept learning. Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society, pp. 304-309, Hillsdale, NJ: Erlbaum, 1995.
- [Towell et al. , 1991] Geoffrey G. Towell, Mark W. Craven & Jude W. Shavlik. Constructive Induction in Knowledge-Based Neural Networks. In L. Birnbaum & G. Collins (Eds.): *Proc. of the 8th Int. Workshop on Machine Learning*, pp. 213-217, San Mateo, CA: Morgan Kaufmann, 1991.

[Wieland & Leighton, 1988] Alexis Wieland & Russel Leighton. Shaping schedules as a method for accelerated learning. *First Ann. INNS Meeting*, p. 231 (abstract), Boston, 1988.