Introduction

Les systèmes biologiques qui nous entourent présentent une incroyable diversité de structures leur permettant d'évoluer et de s'adapter à leur environnement par l'échange, le stockage et le traitement d'information. Autrement formulé, ces systèmes biologiques présentent des capacités de calcul remarquables. Un parangon de système biologique de calcul est sans conteste le cerveau, qui est capable d'exécuter des tâches de calcul et d'apprentissage incroyablement sophistiquées via une multitude de signaux électriques et chimiques circulant entre les neurones et dans les vaisseaux sanguins. L'inspiration biologique occupe ainsi une place de premier rang dans les débuts de la recherche en intelligence artificielle. Les premiers modèles d'apprentissage automatique ont été développés en s'appuyant sur des modèles biologique de neurones, afin de chercher à imiter les capacités d'évolution et d'adaptation à l'origine de la notion d'apprentissage dans les réseaux de neurones biologiques. Le perceptron s'appuyait par exemple sur un modèle simplifié de neurone biologique (McCulloch et Pitts 1990). Les architectures de *Deep Learning* qui en découlent se sont ensuite éloignées de la biologie pour s'extraire des contraintes physiques liées au neurone. Cette approche plus computationnelle a conduit à la création des architectures d'apprentissage performantes que nous connaissons aujourd'hui. Néanmoins, la biologie, par sa diversité de comportements encore incompris, reste une source d'inspiration abondante pour apporter des paradigmes alternatifs ou complémentaires aux modèles d'apprentissage existants. De plus, le comportement du cerveau est loin d'être entièrement compris et modélisé. Les possibilités d'inspiration biologique sont donc constamment en train d'évoluer.

De nombreuses modélisations générales du cortex cérébral, telles que Binzegger et al. 2005 ; Meunier et al. 2009 ; Sporns 2013 ; Betzel et Bassett 2017 proposent que le cortex est une architecture composée de modules auto-organisés. Ces modules communiquent autour des informations sensorielles collectées par l'organisme. Cette communication est réalisée de façon interne, liant des informations sensorielles et abstraites provenant de différentes parties du cortex et à différentes échelles spatiotemporelles. Enfin, bien qu'une hiérarchie de traitement de l'information apparaisse entre ces modules, certains traitant des entrées sensorielles et d'autres des entrées plus abstraites, de nombreux circuits de rétroactions entre les modules semblent présents à différents niveaux de l'architecture. Cette propriété de modularité est partagée par de nombreux systèmes biologiques et artificiels et présente des avantages en termes de réutilisation, de robustesse aux

fautes, de redondance et de traitement local de l'information (Clune et al. 2013). Suivant ce constat, la recherche d'architectures cognitives s'inspirant de l'architecture du cortex cérébral est un enjeu de longue date dans la recherche en apprentissage automatique (Kotseruba et Tsotsos 2018). Il s'agit de développer des réseaux de neurones autonomes, capables de mémoire et de prise de décision de façon non supervisée, qui s'inspirent des architectures modulaires présentes dans le cerveau humain et qui cherchent à imiter certains comportements. Le développement de la robotique et de l'apprentissage incarné appelle également à envisager de telles architectures cognitives, qui sont directement liées à la perception sensorielle. Un robot possède en effet de multiples capteurs, qui peuvent être défaillants, ou qui ne sont pas utilisés dans toutes les tâches que le robot doit effectuer. L'incorporation de mécanismes d'apprentissage au sein de tels agents doit également prendre en compte l'aspect temporel et continu du flux de donnée entrant, ce qui appelle à la conception d'architecture d'apprentissage incluant des boucles sensorimotrices et capables de prise de décision autonomes.

Outre l'inspiration biologique liée à la structure corticale, la construction de systèmes d'apprentissage modulaires découle également d'une motivation computationnelle. Définissons plus précisément ce concept de modularité, qui peut prendre des significations très vastes, de l'informatique à la biologie. Dans une définition générale, un système modulaire est un système composé de sous-systèmes, les modules, dont chacun peut être ajouté ou supprimé sans impacter l'architecture des autres modules. Cette vision générale de la modularité est l'approche classique privilégiée en sciences et ingénierie : pour résoudre un problème, on le décompose en sous-problèmes, puis l'on développe des modules visant à résoudre chacun de ces sous-problèmes. Nous nous plaçons dans une définition plus spécifique de cet aspect modulaire en s'intéressant uniquement à des architectures dont les modules communiquent entre eux de façon locale, sans être supervisés par un processus externe. Cette vision de la modularité se rapproche plus de la vision biologique, dans laquelle aucun processus global ne vient a priori superviser l'organisation des systèmes.

Nous pensons que cette approche modulaire est propice à faire émerger des nouveaux paradigmes d'apprentissage au sein d'architectures neuronales. Le comportement global du système résulte en effet de l'interaction entre les modules et non seulement de la somme des comportements des modules pris individuellement : il s'agit de systèmes complexes. Des exemples de modèles artificiels ont exploré cette idée de modularité. Un exemple ancien en robotique autonome est l'architecture de *sumsumption* de Brooks 1986. Ces travaux construisent une architecture robotique composée de modules comportementaux simples, tels que « marcher », « éviter un objet », mais exploités en architecture par la présence de boucles sensori-motrices. L'interaction de tous ces comportements de base permet au système de réagir de manière autonome à son environnement. Dans cet exemple, les modules ont une structure préétablie. Pour intégrer le contexte d'apprentissage, nous centrons encore notre champ d'intérêt sur des architectures modulaires dont ces modules sont a priori indifférenciés et interchangeables, et vont se spéciali-

ser dans l'architecture au cours de l'apprentissage. En résumé, nous entendons par architecture modulaire d'apprentissage une architecture composée d'une multiplicité de sous-systèmes indifférenciés, interchangeables et évoluant dans le temps. Ils communiquent entre eux par une interface bien définie, et présentent des boucles de rétroaction, leur conférant un aspect dynamique. Cette interaction est traitée localement au sein des modules, sans supervision par un processus extérieur.

Nos travaux s'intéressent à un modèle d'apprentissage initialement inspiré de la biologie : les cartes de Kohonen (Kohonen 1982). Ces modèles sont caractérisés par leur capacité à représenter des données de façon ordonnée sur un espace de dimension plus faible (typiquement une ou deux dimensions). L'algorithme d'apprentissage d'une carte auto-organisatrice suit un principe assez simple. Une carte est composée de vecteurs de l'espace d'entrée (prototypes) positionnés sur une grille de faible dimension. Ils sont initialement distribués aléatoirement dans l'espace d'entrée. L'apprentissage est réalisé en présentant les entrées une à une à la carte, en trouvant leur *Best Matching Unit* qui est le prototype le plus proche de l'entrée, puis en déplaçant ce prototype ainsi que ses voisins dans la grille vers l'entrée. À l'issue de ce processus d'apprentissage, la grille munie des prototypes est dépliée sur l'espace d'entrée. N'importe quel point de l'espace d'entrée peut alors être représenté par une position sur la grille. Cette représentation positionnelle fait des cartes de Kohonen un modèle simplifié l'organisation spatiale observée dans les aires corticales.

La littérature autour des cartes de Kohonen est extrêmement fournie, en témoigne la bibliographie étendue réunissant 7717 travaux entre 1981 et 2005, réunie par Kaski et al. 1998 ; M. Oja et al. 2002 ; Honkela et Kohonen 2009. Toutefois, elle s'est principalement attachée à l'augmentation des performances des cartes sur des applications d'apprentissage automatique et de fouille de données, comme de la compression d'image ou du clustering (Kohonen 2013). Nous pensons que leur inspiration biologique, leurs propriétés d'auto-organisation et de représentation en deux dimensions d'un espace complexe et la simplicité de leurs règles de mise à jour en font des candidates naturelles pour la création d'une architecture modulaire d'apprentissage. D'une part, les cartes auto-organisatrices peuvent être vues comme un modèle très simplifié d'une aire corticale. Leur assemblage en architecture permettrait de pousser cette inspiration biologique au niveau de la structure corticale. D'autre part, elles définissent une représentation en faible dimension de l'espace d'entrée, accessible par les positions dans la carte. D'un point de vue computationnel, cette représentation positionnelle se place comme une information peu coûteuse à échanger au sein d'une architecture.

L'idée d'architecture modulaire de cartes semble donc découler naturellement de la nature même de l'algorithme, et est d'ailleurs formulée par Kohonen dès 1995 :

> « Un objectif à long terme de l'auto-organisation est de créer des systèmes autonomes dont les éléments se contrôlent mutuellement et apprennent les uns des autres. De tels éléments de contrôle peuvent être implémentés par des SOMs spécifiques ; le

> problème principal est alors l'interface, en particulier la mise à l'échelle automatique des signaux d'interconnexion entre les modules et la collecte de signaux pertinents comme interface entre les modules. Nous laisserons cette idée aux recherches futures. » (Traduit de Kohonen 1995)

Depuis, bien que des travaux aient proposé des architectures de carte auto-organisatrices, peu ont effectivement exploré l'aspect topographiquement ordonné et la simplicité des règles mise à jour des poids d'une carte pour les assembler en architectures modulaires comportant des rétroactions : des architectures non-hiérarchiques.

Au vu des propriétés des cartes de Kohonen et de la littérature, nous proposons dans cette thèse de construire une architecture modulaire non-hiérarchique de cartes. L'approche que nous avons privilégiée pour la construction d'une telle architecture est ascendante : nous définissons un modèle de carte pouvant être utilisé en tant que module, puis nous cherchons à comprendre les comportements qui émergent de l'association des modules, afin de guider les améliorations ou applications qui en découlent. L'architecture que nous proposons va dans l'idée d'implémentater des mécanismes généraux liés à la cognition tels que l'apprentissage non-supervisé, autonome, le traitement de données temporelles, l'apprentissage sur le long terme sans oubli catastrophique des données précédentes et la fusion de données multimodales, s'inspirant du traitement multi-sensoriel du cerveau humain. Cette problématique étant extrêmement vaste, nous avons choisi dans cette thèse de nous concentrer sur la tâche particulière d'apprentissage associatif de données multimodales. Il s'agit pour l'architecture d'apprendre des relations existant entre des entrées provenant de différents espaces, en plaçant cet apprentissage de relations à un niveau interne à l'architecture et non en combinant les entrées a priori. Le but est d'apprendre à la fois une représentation des modalités et de leurs relations, tout en gardant une sémantique sur chaque modalité.

*

En résumé, cette thèse vise à répondre à deux problématiques principales entremêlées : (i) développer un modèle d'architecture non-hiérarchique de cartes auto-organisatrices exploitant l'aspect topographiquement ordonné de ce modèle d'apprentissage, et (ii) élaborer une méthodologie expérimentale et des outils permettant de mettre en évidence et évaluer l'apprentissage associatif qui émerge d'une telle architecture.

*

Le manuscrit est organisé de la façon suivante. Le chapitre 1 présente un état de l'art des architectures de cartes auto-organisatrices existant dans la littérature. Ces modèles d'architectures sont issus de plusieurs domaines, de l'apprentissage automatique aux neurosciences computationnelles. Le chapitre propose une revue des modèles principaux en s'attachant à unifier les notations

et leurs désignations afin d'identifier les points communs et différences principales de conception de ces modèles. Cet état de l'art nous permettra de situer le modèle que nous proposons au regard de la littérature existante.

Nous détaillerons au chapitre 2 le modèle d'architecture non-hiérarchiques de cartes auto-organisatrices que nous développons et étudions dans cette thèse, que nous avons appelé CxSOM, pour *Consensus-driven Multi-SOM*. Il s'inscrit dans la continuité de modèles développés dans l'équipe de recherche. Nous définissons un modèle de carte qui peut être assemblé à volonté, de façon modulaire, en architecture non-hiérarchique. Ce modèle utilise la position du Best Matching Unit d'une carte comme seule interface entre les modules, rendant les activités des cartes interdépendantes. Pour gérer les rétroactions, l'apprentissage s'appuie sur une recherche de consensus entre les cartes pour la recherche d'un BMU. Le chapitre 3 est une analyse plus approfondie de la recherche de consensus constituant l'interface entre les cartes afin de valider ce mécanisme en tant qu'algorithme de choix de BMU pour l'apprentissage.

Si notre approche a pour but à long terme de concevoir une architecture comportant de nombreux modules ainsi que des connexions temporelles, nous avons concentré cette thèse sur l'analyse expérimentale des comportements d'apprentissage associatif dans des architectures de deux et trois cartes. Le pari de l'approche ascendante est de faire émerger des nouveaux comportements, des nouveaux mécanismes de calcul ; aussi faut-il pouvoir les mettre en évidence. Nos travaux se sont vite confrontés à une difficulté de visualisation d'une telle architecture de cartes. Cette thèse met l'accent sur une méthodologie d'analyse expérimentale de cette architecture modulaire, ce qui nous permettra d'en tirer des comportements élémentaires qui serviront à poser les bases de la construction d'une architecture plus complexes. Nous introduisons au chapitre 4 cette méthode expérimentale et un cadre de représentation des entrées, et questionnons comment exprimer qu'une architecture de cartes encode les entrées et leurs relations. Nous présentons ensuite au chapitre 5 les comportements élémentaires d'apprentissage associatif observés sur des architectures de deux et trois cartes en une dimension, à partir des représentations que nous avons proposées. Nous présenterons en particulier un comportement de prédiction d'entrée, rendu possible par les rétroactions et la dynamique de recherche du BMU présentes dans le modèle d'architecture. Nous explorons au chapitre 6 des indicateurs numériques d'évaluation de l'apprentissage associatif par l'architecture de cartes, dans le but d'étendre l'analyse du modèle à des architectures plus grandes, qui seraient difficilement représentables graphiquement. Le chapitre 7 étend enfin les mécanismes d'apprentissage que nous avons identifiés à des architectures de cartes en deux dimensions, se plaçant comme une étude préliminaire pour saisir la scalabilité du modèle. Nous conclurons sur les perspectives de développement du modèle CxSOM que mettent en évidence nos travaux.