DeepExpert : vers une Intelligence Artificielle autonome et explicable

Romain Orhand, Anne Jeannin-Girardon, Pierre Parrend, Pierre Collet

DeepExpert : vers une Intelligence Artificielle autonome et explicable

R. Orhand[1,2] A. Jeannin-Girardon[1,2] P. Parrend[1,3] P. Collet[1,2]
[1] Laboratoire Icube - UMR CNRS 7357 – [2] Université de Strasbourg – [3] ECAM Strasbourg-Europe
{rorhand, anne.jeannin, pierre.parrend, pierre.collet}@unistra.fr

Résumé

Pour procurer aux intelligences artificielles une meilleure adaptabilité, notre position est de proposer un couplage selon une approche énactiviste et cognitiviste entre réseaux de neurones profonds et systèmes de classeurs. L'autonomie étant un vecteur d'évolution de cette adaptabilité, une définition est proposée dans un premier temps afin d'appuyer dans un second temps les choix et apports des modèles de perception, de raisonnement et de mémoire employés. Ces modèles, que sont respectivement ces réseaux, ces systèmes et la proposition du concept d'arbre d'apprentissage traçant l'expérience du système couplé, doivent permettre à ces intelligences d'identifier elles-mêmes les problèmes issus de leur environnement et concevoir en toute transparence les solutions les plus adéquates.

Mots Clef

Apprentissage Profond, Systèmes Experts, Arbre d'Apprentissage, Autonomie, Enaction, Explicabilité.

Abstract

To provide artificial intelligence with better adaptability, we suggest a coupling based on an enactivist and cognitivist approach between deep neural networks and learning classifier systems. As autonomy is a vector for the development of adaptability, a definition is first proposed to support later the choices and contributions of perception, reasoning and memory models used. Those models, namely these networks, these systems and a learning tree we design, that keeps track of the experiences of the coupled system, should allow these intelligences to identify, by themselves, problems arising from their environment and design in complete transparency the most suitable solutions.

Keywords

Deep Learning, Expert Systems, Learning Tree, Autonomy, Enaction, Explainability.

1 Introduction

L'intelligence artificielle peut être définie par la capacité d'un système à correctement interpréter des données externes, d'apprendre de celles-ci et d'utiliser leurs apprentissages dans le but d'accomplir des tâches [7].

Dans le but de procurer à ces systèmes une meilleure adaptabilité, une piste intéressante est de développer leur autonomie en s'appuyant sur les fonctions cognitives de l'Homme. Celles-ci sont au moins composées de l'imagination, de la raison et de la mémoire, qu'il est possible de retrouver en partie au sein des ordinateurs, au travers de l'intelligence artificielle [14]. Par ailleurs, deux de ces fonctions cognitives ont été étudiées par Spinoza, selon l'idée que chaque chose se maintienne dans son environnement [15]. Pour ce faire, l'imagination, qui produit des représentations incomplètes, basées sur l'expérience et pouvant être entendue comme la perception, doit servir de support à la raison, afin d'aider ce maintien [9, 15].

Dès lors, une analogie est possible d'une part, entre les réseaux de neurones profonds et la perception et, d'autre part, entre les systèmes de classeurs et la raison : ces réseaux possèdent la capacité d'extraire des représentations intrinsèques aux données et nécessaires à la détection ou classification d'éléments [8], quand ces systèmes sont capables de raisonner à partir de faits en s'appuyant sur une base évolutive de règles et d'apprendre de leurs interactions avec leur environnement [1].

Afin d'étayer cette piste, nous proposons donc, dans ce papier de positionnement, que le concept d'autonomie soit mis en œuvre en couplant, d'une part, des modèles d'apprentissage automatique que sont les réseaux de neurones profonds et, d'autre part, des systèmes experts que sont les systèmes de classeurs. Nous souhaitons de plus donner à ce couplage une mémoire de ses expériences et proposons une nouvelle structure, un arbre d'apprentissage, qui tiendra compte de l'évolution de l'ensemble du système artificiel et lui donnera le droit à l'erreur. Les ordinateurs pourraient ainsi identifier des problèmes et concevoir eux-mêmes les solutions les plus adéquates, tout en exposant le raisonnement élaboré sur les représentations construites.

La notion d'autonomie est d'abord développée en section 2, pour pouvoir expliciter la mise en œuvre du couplage que nous proposons en section 3, avant de conclure dans la section 4.

2 Gagner en autonomie par l'expérience

L'autonomie est définie différemment selon l'interprétation qui est faite de son étymologie grecque (αὐτός et νόμος désignant respectivement *soi-même* et *loi*) ou du contexte dans lequel elle est employée. Elle peut renvoyer vers la faculté à se gouverner soi-même selon ses propres règles,

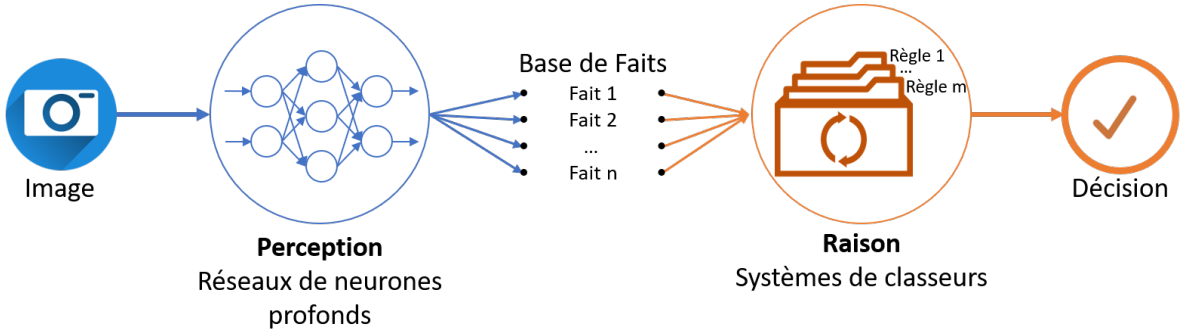

FIGURE 1 – Illustration de DeepExpert : un couplage entre réseaux de neurones profonds et systèmes de classeurs

vers la capacité à être indépendant ou encore vers la génération et le maintien d'une identité [10]. Il n'existe donc pas de définition arrêtée pour l'autonomie et, par extension, de cette faculté pour les ordinateurs.

Cependant, un parallèle est réalisable entre l'autonomie et l'apprentissage automatique, qui désigne la capacité des ordinateurs à apprendre sans être explicitement programmés. Cette définition, attribuée à Samuel [13], souligne le fait qu'un ordinateur se dote de ses propres règles grâce à un apprentissage. Dès lors, nous proposons de définir l'autonomie d'un ordinateur comme sa capacité à réaliser une tâche selon ses propres règles et selon son contexte, en considérant celui-ci du point de vue de l'approche énactiviste [17], puisque l'ordinateur réalise son apprentissage selon ses interactions avec son contexte.

Le couplage proposé fait ainsi appel à l'apprentissage automatique tant dans son modèle de perception, que dans son modèle de raisonnement, ce qui permet de le faire co évoluer avec son environnement. De plus, quand les réseaux de neurones font émerger des représentations de l'environnement propres à l'expérience du système artificiel et à ses capacités perceptives, l'usage des systèmes de classeurs rend possible l'interprétation de ses représentations, en générant des règles intelligibles. Il y a donc constitution d'une interface entre l'Homme et le modèle, favorisant ainsi la création de sens en son sein. Nombre de principes de l'énaction décrits par Varela ou de Loor se retrouvent donc dans ce couplage [4, 17].

Par conséquent, le développement de l'autonomie ne peut s'effectuer sans l'expérience puisqu'elle dépend conjointement de la conception des modèles d'apprentissage, des tâches à réaliser et des interactions avec l'environnement où agit ce modèle. De là, il convient d'expliciter ce système couplant perception, raison et mémoire.

3 Mise en œuvre d'une intelligence artificielle autonome et explicable

L'idée d'une association de réseaux de neurones profonds avec des systèmes de classeurs n'est pas nouvelle : en 2003, Bull proposait les *Neural Classifier Systems* où chaque règle est remplacée par un perceptron multicouche [2]. D'autres modèles ont suivi, tels que les *Neural-Based Learning Classifier Systems* [3] ou les *Spiking Classifiers* [6], mais un des principaux avantages à l'usage des systèmes de classeurs, qui réside en leur explicabilité grâce à la génération de règles intelligibles, est amoindri dans les modèles cités. Récemment, Matsumoto a proposé un système de classeurs basé sur un auto-encodeur profond, le DCAXCSR [11], dont notre proposition se rapproche. Cependant, l'un des inconvénients majeurs de ce modèle est qu'il ne met pas en œuvre un mécanisme de mémoire tel que recherché dans notre approche cognitiviste et énactiviste.

Pour pallier ces inconvénients, le modèle issu du couplage proposé entre la perception et la raison, appelé DeepExpert, est illustré figure 1 au travers d'un simple cas d'usage : en prenant l'exemple d'images issues de l'environnement en entrée, les réseaux de neurones profonds les séparent dans des classes distinctes qui constitueront la base de faits, sur laquelle le système de classeur effectuera, selon sa base de règles, des inférences qui aboutiront en une décision (ou action) induisant un changement dans l'environnement perçu par le système.

Concernant la mémoire, il existe différentes manières de la mettre en œuvre dans les systèmes d'apprentissage, telles que les *Long-Short Term Memory* (LSTM) [5]. Les *Memory Networks*, quant à eux, sont des systèmes de mémoire externe permettant de stocker des informations concernant un certain nombre d'états passés du système [16]. Cependant, ces systèmes ont le désavantage d'être figés : on peut alors citer les *Neural Maps* qui contournent cette limitation en autorisant la mise à jour de la mémoire [12].

Néanmoins, dans les *Neural Maps*, l'usage des réseaux convolutionnels profonds et des LSTM impactent l'interprétation des stratégies employées par l'agent : en effet, les techniques d'apprentissage profond employées obscurcissent le raisonnement effectué par l'agent, ce qui rend impossible de justifier avec exactitude les raisons menant à la réalisation de la tâche. Par ailleurs, les *Neural Maps* utilisées sont conçues par l'Homme : l'agent virtuel ne peut donc réaliser des tâches pour lesquelles il n'a pas été développé.

Sur la base de ce modèle, la mémoire de DeepExpert est construite autour de ce que nous appelons un arbre d'apprentissage, dont le but est de tracer l'évolution de l'ap-

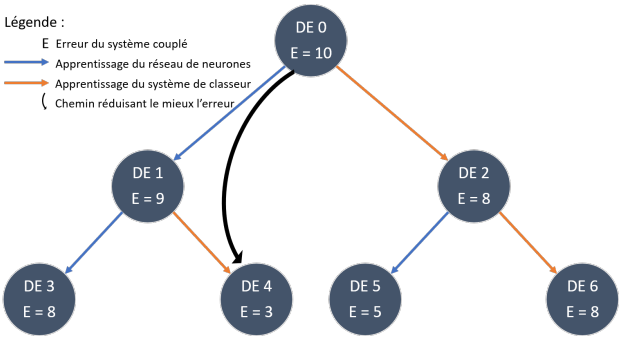

FIGURE 2 – Arbre d'apprentissage de DeepExpert

prentissage de DeepExpert, et de lui donner les outils pour réguler sans intervention humaine son apprentissage.

La figure 2 illustre un tel arbre où chaque nœud représente un état du modèle après apprentissage du réseau de neurones ou du système de classeurs et la racine l'état initial. DE 0 représente donc l'état initial quand le fils gauche DE 1 (respectivement, le fils droit DE 2) représente l'état de DeepExpert après qu'il y ait eu un apprentissage du réseau de neurones (respectivement, du système de classeurs), *etc*. Comme dans tout processus d'apprentissage, le but est de minimiser l'erreur E du système, erreur qui étiquette chaque nœud. La mise en place d'une telle structure favorise alors le développement de DeepExpert sur le long terme en lui octroyant le droit à l'erreur : sur la figure 2, le nœud DE 1 a une erreur de 9 et le nœud DE 2 a une erreur de 8. Pourtant, l'exploration du nœud DE 1 permet d'obtenir une erreur inférieure à celle qui aurait été obtenue en explorant DE 2.

L'arbre d'apprentissage permet donc à notre système d'avoir une mémoire exploratoire. Elle s'intègre également dans le paradigme énactiviste en permettant par exemple la mise en œuvre du concept d'irréversibilité [4] : pour ce faire, il suffit de définir l'un des nœuds en tant que nouvelle racine et de répéter le processus d'apprentissage. De plus, cette irréversibilité nous permet de contrôler la taille de l'arbre, évitant ainsi l'explosion combinatoire.

4 Conclusion

Pour rendre les intelligences artificielles plus adaptables, l'autonomie est employée en tant que vecteur d'évolution de cette adaptabilité, selon une approche énactiviste et cognitiviste : il s'agit donc de la capacité des ordinateurs à réaliser une tâche selon leurs propres règles et leur contexte, et qui émerge des modélisations de la raison, des perceptions et de la mémoire qui sont des fonctions cognitives de l'Homme. Dès lors, les réseaux de neurones profonds, les systèmes de classeurs et la conception d'un arbre d'apprentissage que nous proposons, représentatifs des perceptions, de la raison et de la mémoire, peuvent être utilisés conjointement pour que ces intelligences puissent identifier elles-mêmes les problèmes issus de leur environnement et concevoir les solutions les plus adéquates.

Les perspectives de ce travail sont nombreuses : en premier lieu, l'utilisation d'un système expert devrait permettre de mieux comprendre les décisions prises par DeepExpert, grâce à la génération de règles interprétables. De plus, l'utilisation des systèmes de classeurs permet d'accroître l'autonomie par la génération de règles par évolution artificielle, permettant ainsi à ce système de pouvoir décomposer de manière autonome une tâche en sous-objectifs, après que le concepteur a spécifié l'objectif global à accomplir.

De plus, cette autonomie et cette capacité à générer des sous-objectifs dynamiquement aideraient à mettre en pratique une forme d'apprentissage continu car les nouveaux objectifs permettraient d'ajuster le modèle de perception et ses paramètres, par la nécessité de reconnaître de nouveaux éléments de l'environnement, aidant à la mise en place de nouvelles stratégies pour répondre tant aux sous-objectifs générés, qu'à l'objectif global initial.

Références

[1] C. Buche, C. Septseault, and P. de Loor, Les systèmes de classeurs. Une présentation générale, *TSI. Techniques et Sciences Informatiques*, vol. 25, no. 8-9, p. 963-990, 2006.

[2] L. Bull, On Using Constructivism, in Neural Classifier Systems, *Parallel Problem Solving from Nature - PPSN VII*, p. 558-567, 2002.

[3] H. H. Dam, H. A. Abbass, C. Lokan, and X. Yao, Neural-based learning classifier systems, *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 1, p. 26-39, 2007.

[4] P. de Loor, K. Manac'H, and J. Tisseau, Enaction-Based Artificial Intelligence : Toward Co-evolution with Humans in the Loop, *Minds and Machine*, no. 19, p. 319-343, 2009.

[5] S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, *Neural Computation*, vol. 9, no. 8, p. 1735-1780, 1997.

[6] D. Howard, L. Bull, and P.L. Lanzi, A cognitive architecture based on a learning classifier system with spiking classifiers, *Neural Processing Letters*, vol. 44, no. 1, p. 125-147, 2016.

[7] A. Kaplan, M. Haenlein, Siri, Siri, in my hand : Who's the fairest in the land ? On the interpretations, illustrations, and implications of artificial intelligence, *Business Horizons*, vol. 62, no. 1, p. 15-25, 2019.

[8] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature*, vol. 521, no. 7553, p. 436-444, 2015.

[9] F. Lenoir, *Le miracle Spinoza : Une philosophie pour éclairer notre vie*, Fayard, 2017.

[10] K. Manac'H, *Vers la notion d'agent énactif virtuel : Application à l'approche dynamique évolutionnaire*, Université Européenne de Bretagne, 2011.

[11] K. Matsumoto, R. Takano, T. Tatsumi, H. Sato, T. Kovacs, and K. Takadama, XCSR based on compressed input by deep neural network for high dimensional data, *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, p. 1418-1425, 2018

[12] E. Parisotto, and R. Salakhutdinov, Neural map : Structured memory for deep reinforcement learning, arXiv :1702.08360, 2017.

[13] A. L. Samuel, Some Studies in Machine Learning Using the Game of Checkers, *IBM Journal of Research and Development*, vol. 3, no. 3, p. 210-229, 1959.

[14] M. Serres, Les nouvelles technologies : révolution culturelle et cognitive, *Interstices*, INRIA, 2007.

[15] B. Spinoza, *Ethique*, III, prop. 6.

[16] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, End-to-end memory networks, *Advances in neural information processing systems*, p. 2440-2448, 2015.

[17] F. J. Varela, *Invitation aux sciences cognitives*, Paris, Seuil, 1988.