PSYCHOLOGIE DE L'INTELLIGENCE ARTIFICIELLE : REPERES EPISTEMOLOGIQUES DE L'ANALYSE COGNITIVE DES RESEAUX DE NEURONES

Michael Pichat
Neocognition (Chrysippe R&D), Université de Paris,
Faculté Libre de Philosophie et de Psychologie

Publié sur arXiv le 16 Juillet 2024

Quelle est la « nature » des processus et des contenus cognitifs d'un réseau de neurones artificiels ? Autrement dit, comment « pense » fondamentalement une intelligence artificielle et sous quelle forme résident ses connaissances ? La psychologie de l'intelligence artificielle, prédite en son temps par Asimov (1950), a pour vocation d'étudier cette question. Et cette étude nécessite un niveau neuronal de granularité cognitive, pour ne pas se limiter aux seules résultantes macro-cognitives secondaires (tels les biais cognitifs et culturels) de la cognition synthétique neuronale. Un préalable à l'examen de cette dernière est de clarifier quelques jalons épistémologiques du statut cognitif que nous pouvons accorder à sa phénoménologie.

Que sont les briques cognitives élémentaires d'un réseau de neurones ?

Une couche d'un réseau de neurones opère dans un espace vectoriel dont les dimensions sont associables à des caractéristiques épistémologiques spécifiques à cette couche.

D'un point de vue qualitatif, ces caractéristiques dimensionnelles peuvent être linguistiques (phonémique, phonétique, morphologique, syntaxique, lexicale, sémantique, pragmatique, etc.), visuelles (teinte, saturation, luminosité, contraste, profondeur de dimensionalité, forme, résolution, profondeur de couleur, spectre de couleur, netteté, bruit, texture, contour, composition, échelle, proportion, etc.), auditives (fréquence, intensité, durée, timbre, hauteur, mélodie, rythme, harmonie, bruit, etc.), logiques, contextuelles (position relative, etc.), qualitative, quantitative, etc.. Mais ces caractéristiques peuvent également être de toute autre nature imaginable, que cette nature soit « human like » (i.e.

caractéristiques renvoyant à des catégories de pensée et à des termes que les êtres humains possèdent) ou « alien like » (i.e. caractéristiques renvoyant à des catégories de pensée que les êtres humains ne possèdent pas actuellement) (Bills, 2023).

D'un point de vue typologique, il ne peut exister de classification stable et exhaustive de ces caractéristiques dimensionnelles dans la mesure où elles sont une fonction directe de la nature des données avec lesquelles il est décidé d'alimenter un réseau de neurones donné et des modalités décrétées de son système d'apprentissage, dont le type singulier de feedback qui va lui être administré dans le cas d'un apprentissage (totalement ou partiellement) supervisé. De plus, ces caractéristiques sont le fruit immédiat de l'architecture spécifique allouée au réseau de neurones, de son nombre de paramètres et de leur nature, de la nature de ses opérateurs mathématiques constitutives.

Au sens de la logique formelle, ces caractéristiques dimensionnelles peuvent être associées à des arguments et à des prédicats (propriété, relation, transformation) ou à des combinaisons de ceux-ci. Ces catégories de pensée synthétiques sont des constructions cognitives contingentes et non pas des éléments ontologiques. Elles relèvent d'une infinité de modalités cognitives différentes possibles de segmenter le monde.

Nous reviendrons sur ces deux derniers points après avoir précisé plus avant quelques caractéristiques épistémologiques de la cognition neuronale synthétique.

Quelle est la « nature » de l'activité cognitive opérée par une couche neuronale ?

L'espace vectoriel d'entrée d'une couche neuronale est exprimé dans des caractéristiques dimensionnelles qui sont spécifiques à cette couche. L'embedding d'un input de cette couche est dès lors formaté dans cet espace vectoriel singulier ; tout comme la matrice des poids constitutifs de cette couche est calibrée dans ce même espace vectoriel.

Prémunissons-nous d'emblée d'un piège élémentaire de l'anthropomorphisme cognitif en psychologie de l'intelligence artificielle, en prenant le cas d'un modèle de langage. Une couche neuronale traitant un token entrant, ne raisonne pas sur ce token en tant que tel (à l'instar de notre système cognitif humain, ou en tout cas de notre impression en la matière), ne « voit » pas ce token, mais opère un traitement mathématique sur l'embedding dans lequel est encodé ce token. L'analyse épistémologique de ce traitement mathématique nous enseigne une série de leçons quant à la « nature » de l'activité cognitive neuronale qui est *de facto* portée par ces opérations mathématiques.

A une couche donnée, la fonction d'agrégation de chaque neurone formel opère un traitement mathématique bien déterminé sur l'embedding d'un token entrant. Pour un neurone i donné, ce traitement est généré par le vecteur (horizontal) qui lui est associé et qui contient les poids $W_{i,j}$ qui lui sont propres.

Chacun de ces poids $W_{i,j}$ porte sur une des dimensions catégorielles j de l'espace vectoriel dans lequel sont exprimés les tokens entrants (cf. schéma 1).

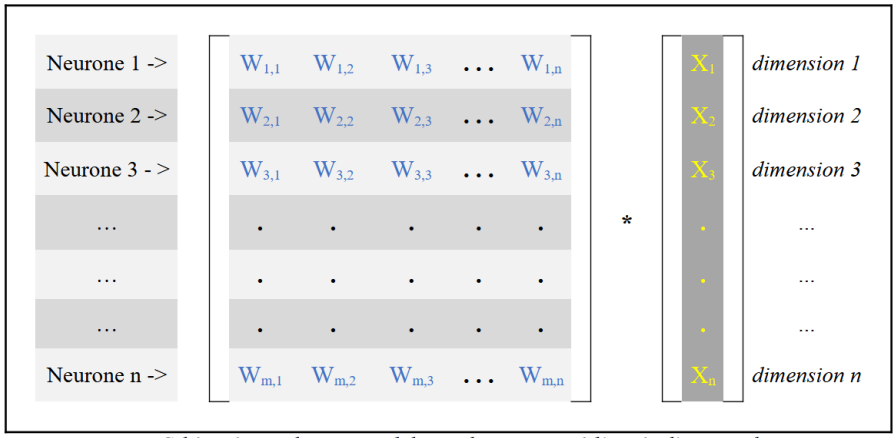

Schéma 1 : produit matriciel des poids neuronaux à l'entrée d'une couche.

Plus précisément, chacun de ces poids $W_{i,j}$ va multiplier une dimension catégorielle j de l'espace vectoriel entrant. Cette multiplication du poids $W_{i,j}$ réalise dès lors une activité cognitive de détermination de l'intensité de focalisation attentionnelle que le neurone i va opérer sur cette dimension j. Autrement dit, un poids est un sélecteur épistémologique qui décide du niveau d'importance (nulle, faible, moyenne, forte, totale) à accorder à une caractéristique dimensionnelle donnée. La valeur résultante de la multiplication opérée indique in fine la combinaison (multiplicative) de deux intensités : (i) l'intensité de possession (X_j) par le token entrant de la dimension catégorielle j, pondérée par l'intensité attentionnelle ($W_{i,j}$) qu'il convient d'accorder à cette intensité de possession catégorielle. Cette valeur exprime donc l'information suivante : quelle est l'importance du niveau de possession (par le token) de la dimension catégorielle ? A ce titre nous pouvons cognitivement la qualifier de résidu épistémologique exprimant un niveau pondéré de possession d'une catégorie épistémologique de segmentation du monde, pour un espace vectoriel d'étape (de segmentation catégorielle progressive) donné (cf. schéma 2).

Puis, dans un second temps, l'ensemble des produits (poids attentionnel x dimension catégorielle) sont additionnés. Cette addition opérationnalise *de facto* une activité de fusion épistémologique pondérée des niveaux de possession, par le token en input, des dimensions catégorielles de l'espace vectoriel d'entrée de la couche neuronale impliquée. Cette fusion épistémologique construit ainsi de toute pièce une nouvelle dimension j' de représentation du token d'entrée, une nouvelle segmentation catégorielle, une nouvelle dimension catégorielle, plus abstraite. Cette fusion cognitive procède par combinaison linéaire de l'ensemble des résidus épistémologiques de ce token initialement formaté dans les dimensions propres à l'espace vectoriel d'entrée j : chaque nouvelle abstraction d'expression du token est donc élaborée par concaténation sélective de la

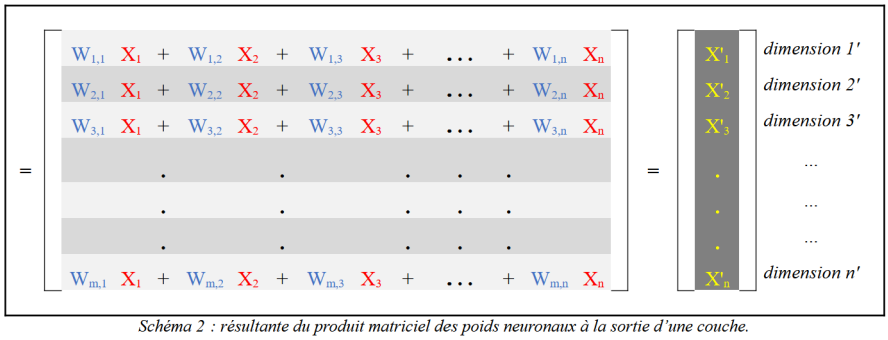

Schéma 2 : résultante du produit matriciel des poids neuronaux à la sortie d'une couche.

totalité des dimensions catégorielles initiales.

La composition mentionnée étant additive, concernant un neurone i donné, plus un token possédera de façon intense les différentes dimensions j de départ sur lesquelles se focalise attentionnellement ce neurone (i.e. pour lesquels les poids $W_{i,j}$de ce neurone ont des valeurs élevées), et plus dès lors la nouvelle abstraction résultante (la dimension j') sera intensément possédée par l'embedding de sortie de ce token. Autrement dit, à l'instar de l'opérateur d'union en logique floue, l'abstraction neuronale artificielle procède par concaténation épistémologique sélective : une nouvelle abstraction de sortie étant la réunion choisie de certaines caractéristiques distribuées dans l'espace (des caractéristiques) de départ, la « signification » d'une possession forte de cette nouvelle abstraction est dès lors celle de la possession intense simultanée de l'ensemble des dimensions superposées de départ dont elle est la résultante de composition sélective.

Ainsi fonctionne, pour partie, le processus cognitif d'abstraction progressive opéré par un réseau de neurones : à l'issue de chaque couche de traitement neuronal, un token est réexprimé dans un nouvel espace vectoriel de sortie dont chaque nouvelle dimension est le fruit d'une « réduction sélective », constitutive de cette nouvelle dimension, de l'ensemble des dimensions d'entrée. Chacune de ces nouvelles dimensions plus abstraites se caractérisant par le mode propre qui a été le sien de combiner sélectivement (i.e. de façon pondérée) la totalité des segments dimensionnels catégoriels de l'espace vectoriel de départ, c'est-à-dire par le fait d'accorder plus ou moins d'importance à chacune de ces catégories dimensionnelles de découpage du monde (des tokens).

Telle est l'activité cognitive portée par les produits matriciels réalisés par les couches successives d'un réseau de neurones : projeter progressivement l'embedding des tokens entrants dans des espaces vectoriels de plus en plus abstraits, chaque nouveau niveau d'abstraction de sortie étant réalisé par fusion épistémologique pondérée des abstractions d'entrée. De couche en couche, les embeddings des tokens sont ainsi reformatés pour exprimer dans des dimensions d'abstraction (d'abstraction d'abstraction etc.) de plus en éthérées et fines.

Quelle est la fonction cognitive des couches neuronales successives ?

La théorie de la conceptualisation développée par Vergnaud (2016) nous offre un cadre précieux pour penser la fonction cognitive des neurones formels d'un réseau de neurones. Selon l'auteur, la forme *princeps* de la connaissance est sa forme opératoire. Cette dernière est constituée de connaissances-en-acte, progressivement fabriquées au fil de l'apprentissage. Ces connaissances sont construites dans l'expérience contingente consistant à devoir agir face à des données de l'environnement. Elles sont nommées « en-acte » dans la mesure où leur fonction n'est pas la théorisation, la formalisation ni même l'explicitation. Leur finalité est en effet avant tout pragmatique : elles sont rattachées à des classes de situations (en réponse desquelles elles ont été construites) et permettent d'en extraire, ou plutôt d'y associer, des caractéristiques fonctionnelles, opératoires dont la prise en compte est déterminante quant à l'efficience de l'activité (cognitive et comportementale) devant être opérée sur ces classes de situations.

Les connaissances-en-acte permettent tout d'abord de découper (segmenter) le flux continu et informe des informations du monde en catégories de pensée élémentaires ; et ainsi de sélectionner, ou plutôt de décréter, des (seules) caractéristiques (issues du monde ou plutôt plaquées sur celui-ci) sur lesquelles il est tenu pour pertinent de focaliser son attention afin de considérer « ce qui est important d'un point de vue pragmatique », « ce qui est source d'efficacité ». Il s'agit de ce que Vergnaud nomme les concepts-en-acte et qui permettent de « lire », de « retenir » ou plutôt de décider d'orienter son regard sur des dimensions particulières du monde (objets, propriétés, relations, etc.) qui sont centrales quant à l'atteinte effective d'un objectif donné. A ce titre, chaque dimension de caractéristique de l'espace vectoriel sur lequel opère un neurone, à l'entrée d'une couche neuronale, possède une fonction cognitive de concept-en-acte. Car ces dimensions, spécifiques à chaque couche, permettent d'identifier les caractéristiques, les concepts analytiques, les abstractions catégorielles, les catégories de pensée dans lesquelles il est fonctionnel de projeter, d'analyser, de catégoriser les informations entrantes de cette couche, afin d'en retenir des aspects fonctionnels relativement à la finalité de la tâche impliquée.

Les connaissances-en-acte, nous indique Vergnaud, permettent également de savoir comment coordonner entre elles les catégories de pensée qui ont été identifiées. Et ainsi de les combiner de façon efficace au sein d'une modélisation cohérente de l'ensemble des critères analytiques qui ont été retenus. Ce processus est réalisé par ce que le chercheur nomme les théorème-en-acte qui sont des micro-théories locales, tenues pour vraies, relatives à la règle par laquelle interagissent les concepts-en-acte. A ce titre, un neurone, ou plus précisément le vecteur des poids définissant un neurone, possède une fonction cognitive de théorème-en-acte. Cela, dans la mesure où ce neurone vecteur-poids est une fonction propositionnelle (f(x,y,z,...)) permettant de composer de façon additive et pondérée les dimensions de l'espace vectoriel d'entrée d'une couche

neuronale donnée ; c'est-à-dire de fusionner de façon sélective certaines de ces dimensions afin d'agir cognitivement selon une modalité particulière sur ces dimensions spécifiques. Chacune de ces fusions cognitives spécifiques (i.e. chacun de ces neurones) ayant pour effet, en termes d'activité cognitive, de générer une nouvelle dimension d'un nouvel espace vectoriel, l'espace vectoriel de sortie de la couche impliquée.

Les nouvelles dimensions catégorielles de l'espace vectoriel de sortie d'une couche, qui sont *de facto* celles de l'espace vectoriel d'entrée de sa couche successrice, ont à leur tour vocation à constituer des concepts-en-acte encore plus fonctionnels qui seront eux-mêmes ensuite combinés par les théorèmes-en-acte neuronaux de cette couche successrice en d'encore nouveaux concepts-en-acte toujours plus abstraits et efficients. Ainsi est la fonction cognitive centrale d'un réseau de neurones en ses couches consécutives : parvenir progressivement, par une série d'étapes d'abstractions conceptuelles croissantes, jusqu'à un niveau de segmentation dans un espace vectoriel catégoriel dont les dimensions ont été suffisamment recombinées sélectivement, affinées et rendues pertinentes pour analyser de façon optimale, au sein de la couche neuronale ultime, les informations reçues par le système neuronal synthétique. La phase d'apprentissage d'un réseau de neurones ayant pour fonction de fabriquer ces étages conceptuels analytiques successifs (i.e. ces connaissances-en-acte) ; et sa phase de fonctionnement ayant alors celle d'exploiter, d'appliquer ces connaissances conceptuelles itératives afin de calculer de façon ultime, les valeurs conceptuelles (i.e. dimensionnelles) de sortie (i.e. l'embedding de sortie) qui produiront (au mieux) la réponse cognitive attendue de ce réseau de neurones.

Les neurones formels ne décryptent pas les propriétés du monde, ils les font émerger en agissant cognitivement sur lui

Les concepts-en-acte neuronaux n'ont pas épistémologiquement de statut ontologique. Ces dimensions catégorielles ne sont pas le fruit d'un illusoire « décodage », par un réseau neuronal formel, de pseudo caractéristiques intrinsèques et préexistantes d'un monde de propriétés qui serait pré-donné et qu'un tel système cognitif synthétique aurait l'habileté de découvrir, de mettre à jour ; habileté cognitive il y a bien, mais de construction enactive au sens de Varela (1988) et non de révélation au sens du réalisme empiriste.

Les caractéristiques dimensionnelles mobilisées par un réseau de neurones sont, nous l'avons déjà mentionné, fonction de l'architecture et de la nature des paramètres qui lui ont été assignés, des données singulières avec lesquelles il a été décidé de l'entrainer ainsi que des opérateurs mathématiques qui lui ont été attribué pour son apprentissage comme pour son fonctionnement. Les concepts-en-acte neuronaux sont dès lors incarnés dans les choix d'opérations mathématiques et de structure de ces opérations qui ont singulièrement présidé à la fabrication de chaque réseau de neurones. Ainsi que le diraient Maturana

(1978) et Von Foerster (2003), les axes vectoriels d'une couche neuronale donnée ne sont pas la résultante d'une « transmission d'information », elles relèvent non pas d'une « copie analogique » mais bien d'une authentique « reconstruction digitale ».

La force d'un réseau de neurones n'est dès lors pas de mettre en lumière les vraies ou bonnes propriétés *per se* des objets du monde ; c'est-à-dire de fabriquer des représentations qui seraient des miroirs fidèles, adéquats de prédicats inhérents au monde. Mais, dans une logique de circularité systémique et constructiviste chère à Varela, de faire émerger des régularités fonctionnelles (en l'occurrence des paramètres du modèle) lors de l'action (cognitive) de celui-ci, avec ses attributs cognitivo-mathématiques qui lui sont propres, sur les objets du monde. Ces régularités traduisant un couplage bilatéral, un modelage mutuel, une articulation structurelle entre le monde et le système cognitif synthétique qui opère sur lui ses opérations de pensée spécifiques.

Prolongements

Nous nous sommes ici bornés à une réflexion épistémologique ayant trait aux fonctions d'agrégation des réseaux de neurones. En négligeant pour le moment les sujets également centraux (i) des fonctions d'activation et de leurs effets topologiques, (ii) des têtes attentionnelles et de leurs impacts contextuels et (iii) des différents types d'architectures neuronales. Ces sujets sont également à questionner dans une dynamique épistémologique.

Plus largement encore, dans une perspective d'hybridation neuro-symbolique (Alshmrany, 2024 ; Sun, 2024) de nos technologies d'intelligence artificielle, il semble nécessaire de s'interroger sur le statut d'une possible articulation directe, à basse granularité, des connaissances-en-acte neuronales synthétiquement générées et des connaissances symboliques formelles humaines. Coordination de connaissances formelles et empiriques permettant de les réguler et de les enrichir mutuellement dans une perspective développementale telle que mentionées par Vygotsky (1934). Coordination apparaissant comme un des prérequis, parmi d'autres issus des neurosciences (Minsky, 1988), à une prochaine génération, radicalement plus évoluée, de systèmes d'intelligence artificielle.

Bibliography

[1] Alshmrany, K. M., Aldughaim, M., Wei, C., Sweet, T., Allmendinger, R., & Cordeiro, L. C. (2024). *FuSeBMC AI: Acceleration of Hybrid Approach through Machine Learning.* arXiv (Cornell University). https://doi.org/10.48550/arxiv.2404.06031

[2] Asimov, I. (1950).*I, robot.* New York : Gnome Press.

[3] Bills, S., Cammarata, N., Mossing, D., Saunders, W., Wu, J., Tillman, H., Gao, L., Goh, G., Sutskever, I., & Leike, J.

(2023). *Language models can explain neurons in language models.* OpenAI. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html

[4] Maturana, H. (1978). *Biology of language: The epistemology of reality..* arXiv (Cornell University). https://doi.org/10.48550/arxiv.2402.06287

[5] Minsky, M. (1988). *The society of mind..* New York : Simon & Schuster.

[6] Sun, R. (2024). *Can a Cognitive Architecture Fundamentally Enhance LLMs? Or Vice Versa?* https://arxiv.org/abs/2401.10444

[7] Varela, F. (1988). *Cognitive Science: A Cartography of Current Ideas.* New York/Leuven: Pergamon Press/Leuven University Press.

[8] Vergnaud, G. (2016). *The Nature of Mathematical Concepts.* In *Learning and Teaching Mathematics* (pp. 5-28). Psychology Press.

[9] Von Foerster, H. (2003). *On constructing a reality. Understanding understanding..* Essays on cybernetics and cognition, 211-227.

[10] Vygotsky, L. S. (1934). *Thought and Language.* MIT Press.