Dossier

Tester nos théories cognitives avec l'Intelligence Artificielle (IA) moderne : le cas du Global Workspace

RUFIN VANRULLEN

CerCo, CNRS UMR 5549, Université de Toulouse, Pavillon Baudot, Hôpital Purpan, 31052 Toulouse Cedex

Comment de nouveaux réseaux de neurones profonds, inspirés de la théorie cognitive du Global Workspace, nous permettent-ils de vérifier l'apport fonctionnel de cette architecture computationnelle ?

La théorie du Global Workspace.

Les neurosciences cognitives regorgent de théories promettant d'expliquer l'émergence de phénomènes complexes tels que la perception, la mémoire, la cognition ou même la conscience. Prenons l'exemple de la théorie du « Global Workspace » proposée par B. Baars pour décrire la cognition et la conscience humaines et animales (1). Un ensemble de « modules » de traitement neuronal spécialisés entrent en compétition pour l'accès à un espace de travail global (le Global Workspace ou GW). Cette compétition est arbitrée par un processus attentionnel tenant compte de l'état du système et de ses objectifs. L'information qui entre dans le Global Workspace est ensuite diffusée à l'ensemble des modules, donnant lieu à une représentation unifiée pour le système entier (c'est le « broadcast »).

Comment savoir si le cerveau fonctionne vraiment de cette manière ? En pratique, on teste plutôt des prédictions expérimentales qui découlent—plus ou moins directement—de la théorie. Il faut d'abord postuler quels composants, structures et processus neuronaux remplissent les différentes fonctions supposées : modules, compétition, workspace, broadcast… C'est ce que propose, par exemple, l'hypothèse du Global *Neuronal* Workspace de S. Dehaene et collègues (2). Il devient ainsi possible de tester expérimentalement les prédictions attenantes, comme dans l'exemple récent d'une collaboration « adversairielle[1] » entre partisans du Global Workspace et ceux d'une théorie concurrente, la théorie de l'Information Intégrée ou IIT (3). Les seuls résultats expérimentaux, pourtant, ne suffisent pas à valider ou invalider l'une ou l'autre théorie : même en cas d'échec, peut-être est-ce la correspondance entre théorie et processus neuronaux qui serait à revoir, plutôt que la théorie elle-même ?

Comment tester *directement* l'idée centrale d'une théorie ? Pour le GW : comment savoir si une architecture cognitive combinant des modules spécialistes, connectés dynamiquement à une représentation centrale partagée, pourrait suffire à engendrer des capacités cognitives avancées ? Il faudrait pouvoir produire artificiellement de telles architectures, afin de les évaluer sur des tâches complexes - une idée techniquement irréalisable, jusqu'à très récemment…

IA et apprentissage profond

Au cours des dernières années, les progrès de l'IA et notamment de l'apprentissage profond (ou « deep learning ») ont constamment redéfini les frontières de ce que l'on pensait être techniquement réalisable. Le deep learning consiste à entrainer des réseaux de neurones artificiels comportant des dizaines de couches, des millions de neurones et jusqu'à plusieurs milliards de connexions, afin de résoudre une tâche prédéfinie. Classifier une image, reconnaître un visage ou une voix,

[1] *Ce terme est emprunté aux GANs, les réseaux génératifs adversairiels qui améliorent leurs performances en opposant deux sous-réseaux « adversaires », entraînés conjointement.*

maîtriser un jeu vidéo ou de stratégie, mais aussi générer de toutes pièces une image, une vidéo réaliste ou un morceau de musique, ou simplement répondre à des questions et requêtes farfelues en langage naturel—combien de ces capacités aurait-on prédit que l'IA pourrait atteindre, il y a seulement 10 ans ?

Une conséquence de ces avancées de l'IA pour les neurosciences est la possibilité inédite de déployer nos théories cognitives sur des systèmes artificiels à grande échelle, et ainsi les mettre à l'épreuve (4).

Une implémentation IA du Global Workspace

C'est ce que nous avons entrepris de faire pour la théorie du Global Workspace (5). La procédure consiste à interconnecter différents réseaux préexistants qui jouent le rôle des modules spécialistes. Ceux-ci peuvent traiter des entrées visuelles, audio, ou du langage, mais également produire des images, du son, du texte, voire même contrôler les actions d'un robot. Un système attentionnel dédié décide du ou des modules qui pourront accéder au GW à chaque instant, en fonction de l'état du système, de ses objectifs et du contenu actuel de chaque module. Lorsqu'un module est sélectionné par l'attention, il implante son contenu dans le GW ; pour cela, un encodeur spécifique de chaque module convertit son activité en une représentation commune, une sorte de *lingua franca* ou langue véhiculaire du Global Workspace. Il s'ensuit une diffusion de cette représentation commune à l'ensemble des modules, appelée « broadcast » ; pour cela, chaque module dispose également d'un décodeur pour retranscrire le langage commun du GW dans son langage propre. On voit dans cette métaphore des langages propres et communs que l'opération du GW consiste principalement en une forme de traduction neuronale. Ainsi on peut faire appel à des outils d'apprentissage non supervisé [2], traditionnellement utilisés en traitement du langage (6) : pour entraîner les encodeurs et décodeurs qui permettent d'échanger l'information entre modules et GW, on s'appuie sur un cycle de traduction suivie de la traduction inverse. S'assurer qu'un tel cycle reste cohérent revient à aligner entre eux les espaces de représentation des modules et du GW, sans pour autant connaître la « vraie » traduction (si tant est qu'elle existe). Ceci implique que l'entraînement d'une architecture de Global Workspace nécessitera moins de données et moins de supervision explicite qu'un apprentissage classique (entièrement supervisé).

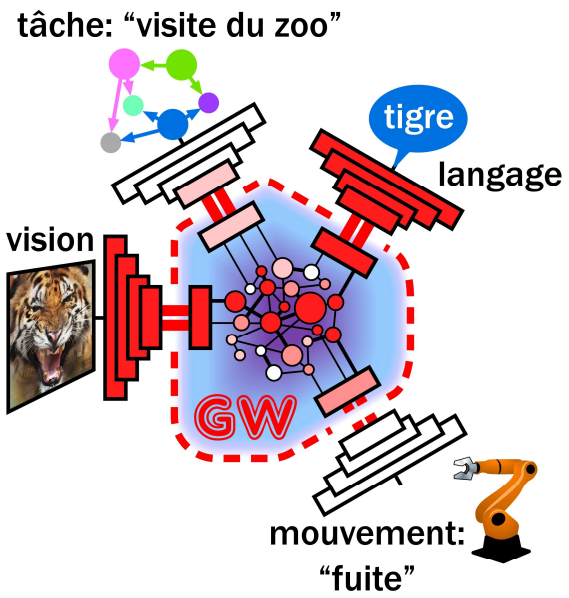

Figure 1: *Schéma illustratif du Global Workspace (GW) dans nos implémentations IA. Dans cet exemple, 4 modules spécialisés sont connectables au GW (en bleu). Ces modules pré-entraînés pour différentes fonctions (classification d'images, génération de texte, contrôle d'actions robotiques ou représentations de tâches) sont capables d'opérer indépendamment du GW. Cependant, un système attentionnel (non représenté ici) peut choisir de connecter chaque module au GW. Dans ce cas, l'information traitée par le module accède au GW, et ce faisant, devient elle-même accessible à l'ensemble des autres modules via la diffusion ou « broadcast ». Même un module qui n'est pour l'instant pas connecté au GW (dans cet exemple, le module de mouvement) a ainsi connaissance de la présence du tigre, traduit dans un format pertinent pour ce module (ici, la préparation d'une réponse motrice de fuite). Si le module était connecté, cette connaissance donnerait lieu à une réponse (de fuite) effective. Ainsi, le système entier a la capacité de contrôler son comportement et de l'adapter à la situation de manière flexible.*

Premiers résultats

En effet, nous avons pu confirmer l'intérêt d'apprendre une représentation multimodale via le GW, dans des scénarios simplifiés où seuls deux modules de représentation (vision, langage) sont connectés au GW (7,8). Par rapport à un entraînement supervisé classique, l'utilisation des cycles de traduction (reflétant le phénomène de broadcast) permet de diviser par un facteur de 5 à 10 le nombre de paires d'exemples [image, texte] requis, pour un niveau de performance équivalent (7). De plus, la représentation multimodale apprise au sein du GW permet d'atteindre de meilleures

[2] *L'apprentissage non supervisé s'oppose à l'apprentissage supervisé, où l'on indique directement au système la sortie désirée pour chaque entrée.*

performances quand on l'utilise ensuite pour des tâches de classification, par rapport à un système multimodal conventionnel tel que le célèbre modèle CLIP d'OpenAI (9) (CLIP : Contrastive Language-Image Pretraining).

De même, nous avons pu utiliser la représentation du GW pour entrainer un agent d'apprentissage par renforcement (« reinforcement learning ») à contrôler un robot dans un environnement industriel simulé (8). Grâce à sa méthode d'entrainement basée sur le broadcast, le système GW s'avère robuste à des changements de modalité d'entrée : par exemple, si le système apprend initialement à contrôler le robot sur la base d'informations numériques décrivant sa position dans l'environnement, il pourra ensuite généraliser la stratégie apprise lorsqu'on lui donnera des images en entrée au lieu des informations numériques (on parle de transfert « zero-shot »). Un système équivalent mais entraîné à la manière de CLIP (9) ne démontre pas ces capacités de généralisation et de robustesse.

Perspectives

Ces premiers éléments contribuent à établir que l'architecture GW apporte bien un avantage fonctionnel aux systèmes cognitifs artificiels. Cependant, de tels résultats nécessiteront d'être confirmés et étendus dans plusieurs directions. Tout d'abord, il faut considérer d'autres modalités d'entrée/sortie que l'image et le texte : par exemple, l'audio, la parole, et l'action ; mais aussi des processus internes comme l'émotion, la motivation, la mémoire… Apprendre à aligner ces diverses modalités via le GW (grâce aux cycles de traduction et au broadcast) reviendrait à appréhender les analogies qui existent naturellement entre elles et, ce faisant, à ancrer les représentations du monde dans leurs manifestations internes à l'agent ; et, symétriquement, à ancrer les représentations symboliques de l'agent dans un référentiel extérieur, i.e. dans l'environnement. On peut penser que c'est cet ancrage (ou « grounding ») qui fait défaut aux systèmes d'IA actuels, même les plus performants, comme les grands modèles de langues (LLMs ou Large Language Models). Ainsi, la modélisation du GW pourrait aussi avoir des conséquences pratiques en IA, en rapprochant les LLMs de la cognition humaine et en réduisant leurs erreurs invraisemblables ou « hallucinations ».

Un autre axe de recherche est l'exploration du mécanisme de sélection attentionnelle. Un module peut être sélectionné parce qu'il porte une information intrinsèquement saillante, ou parce qu'il semble bénéfique à un instant donné, ou encore, parce qu'il participe à une séquence d'opérations planifiée pour résoudre un problème complexe. Dans tous ces cas, il faut pouvoir entrainer (par apprentissage profond) un contrôleur attentionnel qui prendra correctement en charge cette sélection. Nous développons déjà des versions d'un tel contrôleur permettant de sélectionner parmi plusieurs modalités celle qui porte l'information la moins bruitée, ou encore de recruter itérativement un module d'incrément numérique pour résoudre une tâche d'addition. À terme, cette architecture générique devrait pouvoir adapter le comportement et les réponses de l'IA à sa situation et ses objectifs, à des horizons temporels plus ou moins lointains, c'est-à-dire manifester une aptitude cognitive de « système 2 » (par opposition à une cognition plus automatique et réflexive de « système 1 »).

Améliorer les capacités de l'IA est une manière claire de démontrer la viabilité de la théorie du Global Workspace. Une autre manière serait de vérifier que l'activité et le comportement de nos systèmes artificiels correspondent bien à ce qu'il se passe dans le cerveau. Nous collectons actuellement des données d'IRMf dans des conditions multimodales (image+texte) afin de comparer l'activité cérébrale à celle du GW (ainsi qu'à d'autres systèmes d'IA multimodale comme CLIP). Au niveau comportemental, nous tentons d'évaluer dans quelles conditions l'apprentissage multimodal du GW favoriserait l'émergence d'associations cross-modales non véridiques, ressemblant à des formes de synesthésie. Ces travaux pourraient engendrer des prédictions expérimentales, vérifiables par les Neurosciences.

Enfin, cette ligne de recherche pose également des questions philosophiques et éthiques non négligeables, dès lors qu'on se rappelle que le Global Workspace n'est pas uniquement une théorie de la cognition, mais est aussi considérée aujourd'hui comme une théorie majeure de la conscience. Doit-on penser qu'une implémentation IA donnerait lieu à une réelle expérience phénoménologique artificielle—probablement pas dans les systèmes rudimentaires que nous avons présentés ici, mais peut-être chez leurs descendants ? La question mérite considération (10).

rufin.vanrullen@cnrs.fr

Références

(1) Baars, B. J. (1993). *A cognitive theory of consciousness,* Cambridge University Press.

(2) Mashour G. A. et al. (2020). *Neuron*, 105(5), 776-798.

(3) Melloni L. et al. (2023) *PLoS ONE*, 18(2), e0268577.

(4) Hassabis D. et al. (2017) *Neuron*, 95(2), 245-258.

(5) VanRullen R., & Kanai R. (2021). *Trends Neurosci*, 44(9), 692-704.

(6) Artetxe M. et al. (2018). *Unsupervised neural machine translation. 6th International Conference on Learning Representations*, ICLR 2018.

(7) Devillers B. et al. (2025) *IEEE Trans Neural Netw Learn Syst*, 36(5):7843-7857. doi: 10.1109/TNNLS.2024.3416701

(8) Maytié L. et al. (2024). *Reinforcement Learning Journal* 3 (170), 1410-1426

(9) Radford A. et al. (2021). *Proceedings of the 38th International Conference on Machine Learning,* PMLR 139:8748-8763.

(10) Butlin P. et al. (2023) *arXiv preprint arXiv*:2308.08708.