NICOLAS GAMBARDELLA
8602C
DIRECTEUR DE RECHERCHE
AU CNRS

L'INTELLIGENCE ARTIFICIELLE

L'IA, COMMENT ÇA MARCHE ?

DES SYSTÈMES IMITANT LE CERVEAU

Le premier contact de la plupart des gens avec l'IA se fait via *ChatGPT* et consort ou les publicités de plus en plus ciblées dont nous sommes assaillis. Mais l'IA au final, c'est quoi ? Qu'y a-t-il sous le capot ?

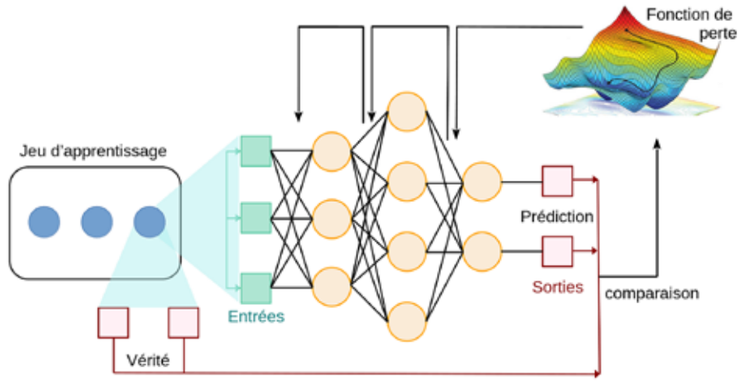
D'abord, il faut définir des termes souvent utilisés de manière interchangeables, mais qui recouvrent des réalités différentes. On appelle "intelligence artificielle" l'ensemble du domaine s'intéressant aux machines ou aux programmes informatiques capable de cognition (le terme "intelligence" étant lui-même controversé). Au sein de ce domaine, on distingue celui de l'apprentissage automatique (*Machine Learning* en anglais), où les outils apprennent les règles à partir de jeux de données qu'on leur fournit plutôt que de les recevoir de leur concepteur a priori (comme les systèmes experts). Au sein de l'apprentissage automatique, un type d'algorithme particulier, fondé sur une structure mathématique appelée neurone artificiel, forme ce que l'on appelle désormais l'apprentissage profond (*Deep learning* en anglais).

Les réseaux de neurones

Un neurone artificiel est juste une fonction mathématique qui fait la somme des entrées venant d'autres neurones, pondérées par des "poids synaptiques" qui vont être appris lors de l'entraînement du modèle. Puis ces fonctions produisent une sortie unique après transformation qui est envoyée vers d'autres neurones. Il existe un grand nombre de fonctions possibles, la plus simple étant l'identité (qui renvoie juste ce qu'elle reçoit) ou bien des fonctions non-linéaires, comme les fonctions en "S" (qui transforment une entrée de $-\infty/+\infty$ en 0/1 ou $-1/+1$). La magie vient de ce que l'on peut créer des couches de ces neurones, et empiler ces couches. De ce fait, n'importe quelle équation mathématique peut être représentée par un réseau de neurones artificiels de même que n'importe quelle courbe dans un nombre quelconque de dimensions. Il existe de nombreuses architectures de réseaux de neurones, depuis les perceptrons multi-couches des débuts jusqu'aux grands modèles de langage d'aujourd'hui, en passant par les réseaux de neurones à convolution pour la reconnaissance d'images. Ces réseaux de neurones peuvent comprendre une poignée de paramètres ou bien des centaines de milliards comme *ChatGPT*.

L'apprentissage

L'apprentissage peut être non-supervisé, c'est-à-dire qu'on laisse les programmes apprendre directement des données, par exemple pour grouper ensemble des objets ayant les mêmes caractéristiques. Cela peut être des images, à savoir des listes de pixels, ou bien des patients, représentés par des données cliniques. Un exemple extrême d'IA non-supervisée est le système *AlphaGo* qui a appris à jouer au Go tout seul, en observant des parties existantes, puis dans ses dernières versions en jouant contre lui-même, sans avoir besoin de jeux de données d'apprentissage. C'est ce que l'on appelle l'apprentissage auto-supervisé. *AlphaGo* a battu le meilleur joueur du monde en 2016. Les auto-encodeurs sont un autre exemple d'apprentissage non-supervisé. Ces systèmes apprennent à reproduire leurs entrées, par exemple des images de chats, à l'identique. On peut dès lors supprimer la partie encodeur pour ne garder que la partie générative (le décodeur). On obtient un très bon générateur d'images de chat.



Représentation simplifiée du mécanisme d'apprentissage d'un réseau de neurones, ici un perceptron à trois couches

À l'opposé, on trouve l'apprentissage supervisé. On va dire au modèle ce qu'il doit prédire, reconnaitre ou produire. Pour cela, on a besoin de grands ensembles de données bien étiquetées. L'existence du jeu de données *ImageNet*, avec ses 14 millions d'images soigneusement étiquetées, a été un élément crucial dans la création des CNN profonds de reconnaissance d'images. La sortie des modèles peut être une valeur chiffrée (e.g., un risque), une classification (e.g, reconnaitre un chat d'un chien), ou un objet (un texte, une image, des sons).

Mais comment le modèle apprend-il ? La clé est la comparaison entre les sorties effectives et attendues. Cette comparaison va fournir une fonction de coût que l'on va tenter de diminuer en changeant les paramètres du modèle (les poids synaptiques) pour que les sorties se rapprochent de ce que l'on attend (le *ground truth* en anglais). L'algorithme clé qui a rendu possible cet apprentissage est la rétropropagation du gradient, qui va modifier les poids en partant des dernières couches, proches des sorties du modèle, puis progressivement remonter vers les entrées[1]. La taille des ensembles d'apprentissage doit être proportionnelle au nombre de paramètres dont la valeur doit être apprise. Par exemple, on pense que *ChatGPT* v4 a appris avec plusieurs centaines de milliards de mots, réunis en quelques dizaines de milliards de portions de texte. Que fait *ChatGPT* en réalité ? Est-ce qu'il comprend ? Est-ce qu'il raisonne ? Non, ChatGPT ne réfléchit pas aux réponses ; il ne sait même pas ce qu'est une question. La seule chose que fait *ChatGPT* est proposer le mot suivant, puis le suivant, puis le suivant, etc., le tout en fonction de tous les mots de la conversation. Et pour qu'il ait appris quel mot prédire, on lui a présenté des séries de mots dont on avait masqué certains, en lui demandant de les prédire.

Plongement et attention

Une notion fondamentale en intelligence artificielle est celle de plongement (*embedding* en anglais), où l'on va transporter les données d'un espace de départ, par exemple l'espace des mots du dictionnaire français, où le mot "femme" a une coordonnée de 1 sur l'axe "femme" et 0 sur tous les autres, dans un nouvel espace (l'espace latent) où le mot "femme" sera plus proche de "mère" que de "fourchette" ou de "père". De plus, dans ce nouvel espace, on pourra passer de "femme" à "père" en additionnant les trajets "femme-homme" et "homme-père" (ou "femme-mère" et "mère-père"). De la même façon, un réseau de neurones à convolution transforme les images en une liste de coordonnées dans un espace où les images d'éléphants sont plus proches les unes des autres que des images de chats... ou de fourchettes. Les grands modèles de langages comme BERT transforment les phrases en entrées via plusieurs plongements, correspondant à la position du mot dans le dictionnaire, sa position dans la phrase et la position de la phrase dans le texte. Ainsi, les mots "fourchette" en sujet et en complément d'objet seront encodés différemment.

Finalement, nous voila rendus au concept qui a tout changé il y a une dizaine d'années, en particulier lorsqu'il a été utilisé avec l'architecture dite des "transformers", à savoir l'attention. Les modèles dits "récurrents", qui étaient par exemple utilisés pour prédire l'évolution des cours en bourse, avaient déjà introduits la notion de mémoire, mais cette mémoire disparaissait avec la longueur des entrées (lors de l'apprentissage, du fait d'un phénomène appelé "dissipation du gradient", quand venait le moment de traiter la fin d'un paragraphe le modèle avait oublié le début). L'idée centrale de l'attention est que les entrées successives ne soient plus indépendantes les unes des autres. Les modèles apprennent – sur la base de milliards d'exemples – quelles sont les entrées qui vont affecter la façon dont ils vont prendre en compte une entrée donnée ; en d'autres termes quelle est l'*attention* qu'une entrée donnée va porter aux autres entrées. Un modèle de traduction saura dès lors qu'un "avocat" doit être traduit en *lawyer* s'il est accompagné de "tribunal" et de "plaidoirie" mais en *avocado* s'il est accompagné d'"huile" et de "salade". Ces tables d'attention sont absolument énormes, et croissent généralement[2] avec le carré de la taille des séquences d'entrée. Ce qui signifie un nombre de calculs faramineux durant les phases d'entrainement. Les processeurs graphiques (GPU) ont changé la donne en matière d'efficacité. Mais le nombre d'opération nécessaire les rend tout de même gourmands. Certaines études avancent que l'entraînement de *ChatGPT4* a nécessité plus de 1025 calculs élémentaires, c'est-à-dire 1 suivi de 25 zéros. Les crypto-monnaies et l'IA représentent 2 % de la consommation d'électricité mondiale, cette dernière étant en croissance rapide. ■

1. Tout cela est très simplifié. Les versions modernes de ChatGPT sont très complexes. Elles n'utilisent même pas notre invite, mais utilisent cette dernière pour faire des recherches sur internet et proposer au modèle une invite contextuelle et améliorée. Puis elles vérifient les réponses à l'aide d'algorithmes avancés (par exemple en mathématique pour corriger les réponses).
2. Cela est en train de changer. L'auteur utilise de nouveaux modèles apparus en 2023, les State Space Model dont les demandes croissent de manière linéaire...