**Comment fonctionne ChatGPT ? Décrypter son nom pour comprendre les modèles de langage**

ALEXANDRE Frédéric, Centre Inria de l'université de Bordeaux, CNRS, Bordeaux INP

On voit passer beaucoup d'avis sur des Intelligences Artificielles comme ChatGPT, sur leur dangerosité, leur niveau d'intelligence voire leur humanité. Mais finalement, que sait-on de ChatGPT ? Que c'est un réseau de neurones artificiels composé de milliards de paramètres ; Qu'il est capable de tenir une discussion de haut niveau ; mais aussi qu'il peut tomber dans des pièges grossiers tendus par des internautes facétieux. Bref, on nous parle beaucoup de lui mais on en sait finalement très peu sur son fonctionnement. Donc, avant d'en débattre, je vous propose de présenter les mécanismes principaux sur lesquels ChatGPT repose, ce qui nous permettra de comprendre que, si le résultat est parfois impressionnant, ses mécanismes élémentaires sont souvent astucieux mais pas vraiment nouveaux. Pour ce faire, passons en revue les différents termes du sigle ChatGPT.

**T comme Transformer**

Un Transformer est un réseau de neurones utilisé principalement comme modèle de langage et qui bénéficie du même algorithme d'apprentissage que les réseaux profonds (*Deep Networks*), algorithme qui a donc déjà fait ses preuves pour l'entrainement de grosses architectures. Il bénéficie également de deux caractéristiques éprouvées. D'une part, en tant que modèle de langage, il manipule des séquences de mots, qu'il code avec des techniques dites de [plongement lexical](https://fr.wikipedia.org/wiki/Plongement_lexical), que l'on présente dans le [texte](« Comment ChatGPT comprend-il les mots qu'il utilise ? »).

D'autre part, l'autre caractéristique intéressante concerne la façon dont l'aspect séquentiel des mots est traité. Il s'agit ici d'un problème majeur car cette séquentialité est à prendre en considération (pour interpréter le sens de certains mots dans le contexte plus général de la phrase) et les techniques proposées sont souvent très coûteuses en temps de calcul et relativement peu efficaces. Ici aussi, la technique proposée par les Transformers privilégie une approche numérique et statistique, simple à calculer massivement et très efficace. Il s'agit d'une technique dite attentionnelle qui consiste, pour interpréter le sens de chaque mot, de se demander à quelles parties de la phrase il faut faire attention pour associer un mot à son contexte. Le texte mentionné plus haut explique qu'avec le plongement lexical, chaque mot pouvait être remplacé par un descriptif numérique (un vecteur) de grande dimension. L'idée avec l'approche attentionnelle est d'apprendre comment chacun de ces vecteurs peut également être influencé par le vecteur (le descriptif) de certains autres mots de la phrase, ce qui permet d'accorder un mot ou de remplacer un pronom par le mot de la phrase qu'il représente. Et ici aussi, comme les textes sont, comme les humains, très prévisibles, il est impressionnant de constater comment ce type d'approches statistiques appliquées à des grands corpus permet des interprétations de qualité. A titre d'illustration, voyez à quel point vous êtes capables de lire une BD des [Schtroumpfs](https://fr.wikipedia.org/wiki/Les_Schtroumpfs) et de remplacer chaque 'schtroumpf' par un mot issu de l'analyse attentionnelle des autres mots.

**G comme Génératif**

Ce terme renvoie au fait que ChatGPT est capable de générer du langage : on lui explique un problème, on lui pose une question et, ayant assimilé cette interpellation, il nous répond avec du langage. C'est pour ça qu'on l'appelle 'modèle de langage', car pour ce faire, il doit voir appris un tel modèle. Ici aussi, la possibilité d'apprendre un modèle génératif avec un réseau de neurones date de plus de trente ans et a été décrite sous la forme de modèles [d'auto-encodeurs](https://fr.wikipedia.org/wiki/Auto-encodeur), aussi appelés réseaux diabolo, faisant référence à la forme de ces réseaux. Prenons un exemple simple (illustré figure 1) avec un réseau ayant une couche d'entrée et une couche de sortie de taille identique et une couche cachée de très petite taille.

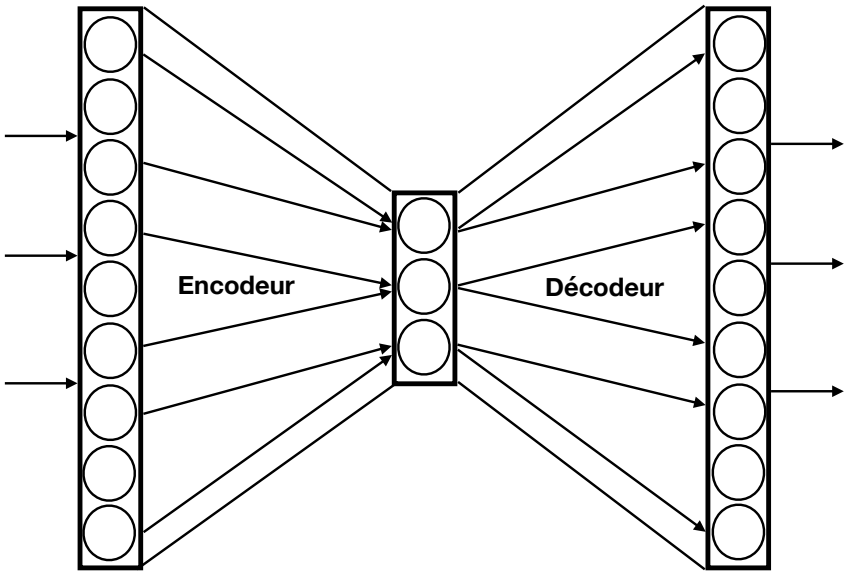


Figure 1 : un auto-encodeur extrait les variables latentes et permet de créer un modèle génératif

On entraine le réseau en lui présentant des phénomènes similaires (par exemple des photos de visages) sur la couche d'entrée et en le supervisant pour qu'il reconstruise la même chose (le même visage) sur la couche de sortie. C'est une tâche assez peu intéressante, sauf si on considère que, pour pouvoir faire cette reconstruction, le réseau doit passer par un codage plus compact dans la couche cachée. C'est l'obtention de ce codage compact qui sera la fonction intéressante de ce réseau (cette partie s'appelle l'encodeur), mais aussi sa capacité à reconstruire la forme d'origine à partir du codage compact (cette partie s'appelle le décodeur). Des études mathématiques ont analysé les caractéristiques de ce codage compact (éliminant les détails superflus et rendant explicites les variables cachées principales du phénomène considéré, aussi appelées variables latentes) et d'autres ont étudié comment il était possible de stimuler la couche cachée (de choisir certaines variables latentes) pour obtenir à travers le décodeur des nouveaux exemplaires du phénomène considéré. C'est par exemple en suivant ce procédé (pour des réseaux de grande taille comportant d'autres couches cachées intermédiaires avant et après la couche cachée

centrale de petite taille) que l'on est capable de créer des [deepfakes](https://fr.wikipedia.org/wiki/Deepfake), c'est-à-dire des trucages très réalistes.

Si, au lieu de photos, on souhaite maintenant considérer des séquences (des vidéos ou des phrases), il faut en plus savoir prendre en compte l'aspect séquentiel du flux d'entrée. Ceci peut être obtenu avec un phénomène attentionnel comme décrit plus haut, qui permet d'apprendre à produire chaque élément de la séquence en y ajoutant un contexte, représentant certains éléments précédents de la séquence, auxquels on aura appris à prêter attention. La possibilité d'utiliser un simple mécanisme attentionnel pour traiter l'aspect séquentiel des entrées a été un constat majeur dans la mise au point des Transformers (« Vous n'avez besoin que d'attention » titrait la publication correspondante; *« Attention is all you need »*), car auparavant les méthodes privilégiées utilisaient des réseaux plus complexes, dits récurrents, dont l'apprentissage reste très lent et imparfait ; de plus ce mécanisme attentionnel se parallélise très bien (les mécanismes alternatifs sont séquentiels !), ce qui accélère d'autant plus cette approche attentionnelle.

**P comme Pretrained**

Les mécanismes décrits plus haut constituent l'essentiel des méthodes utilisées pour construire un Transformer et si beaucoup ont été surpris par leur efficacité, c'est que cette dernière n'est pas seulement due à la puissance de ces méthodes, mais aussi (et surtout ?) à la taille de ces réseaux et des connaissances qu'ils ingurgitent pour s'entrainer. Les détails chiffrés sont difficiles à obtenir, mais on entend régulièrement parler pour des Transformers de milliards de paramètres (de poids dans les réseaux de neurones) ; pour être plus efficaces, plusieurs mécanismes attentionnels (jusqu'à cent) sont construits en parallèles pour mieux explorer les possibles (on parle d'attention « multi-tête »), on peut avoir une succession d'une dizaine d'encodeurs et de décodeurs, etc. Rappelons que l'algorithme d'apprentissage des Deep Networks est générique et s'applique aussi bien quelle que soit la profondeur (et la largeur) des réseaux ; il suffit juste d'avoir assez d'exemples pour entrainer pour ces poids, ce qui renvoie à une autre caractéristique démesurée de ces réseaux, leur corpus d'apprentissage. Ici aussi, peu d'informations officielles, mais il semble que des pans entiers d'internet soient aspirés pour participer à l'entrainement de ces modèles de langages, en particulier l'ensemble de Wikipédia, les quelques millions de livres que l'on trouve sur internet (dont des versions traduites par des humains sont très utiles pour préparer des Transformers de traduction), mais aussi très probablement les textes que l'on peut trouver sur nos réseaux sociaux favoris. Cet entrainement massif se déroule hors ligne, peut durer des semaines et utiliser des ressources calculatoires et énergétiques démesurées (chiffrées à plusieurs millions de dollars, sans parler des aspects environnementaux).

**Chat comme bavarder**

Nous sommes maintenant en meilleure position pour présenter ChatGPT : il s'agit d'un agent conversationnel, bâti sur un modèle de langage, GPT, qui est un Transformer Génératif Pré-entrainé. Les analyses statistiques (avec approches attentionnelles) des très grands corpus utilisés permettent de créer des séquences de mots ayant une syntaxe de très bonne qualité. Les techniques de plongement lexical offrent des propriétés de proximité sémantique qui donnent des phrases dont le sens est souvent satisfaisant. Outre cette capacité à savoir

générer du langage de bonne qualité, un agent conversationnel doit aussi savoir converser, c'est-à-dire analyser les questions qu'on lui pose et y apporter des réponses pertinentes (ou détecter les pièges pour les éviter). C'est ce qui a été entrepris par une autre phase d'apprentissage hors-ligne, avec un modèle appelé InstructGPT, qui a nécessité la participation d'humains qui jouaient le jeu de faire l'agent conversationnel ou de pointer des sujets à éviter (en disant comment les éviter). Il s'agit d'apprentissage dit par renforcement qui permet de sélectionner des réponses selon les valeurs qu'on leur donne ; c'est une sorte de semi-supervision où les humains disent ce qu'ils auraient aimé entendre (ou pas).

Ayant une meilleure compréhension des mécanismes de ChatGPT, on pourra discuter ailleurs plus en profondeur de sa possible dangerosité ou de sa similarité avec l'humain mais on peut conclure rapidement ici que ce n'est pas la peine de comprendre un sujet pour savoir en parler avec éloquence, sans donner forcément de garantie sur la qualité de ses réponses (mais des humains aussi savent faire ça…).