Beau parleur comme une IA

THE CONVERSATION, 26 décembre 2022, par Fabian Suchanek, et Gaël Varoquaux.

Fabian Suchanek est Professeur en informatique, Télécom Paris – Institut Mines-Télécom
Gaël Varoquaux est Directeur de recherche en intelligence artificielle et applications en santé, Inria.

Les intelligences artificielles apprennent à parler grâce aux « modèles de langage ». Les modèles les plus simples permettent la fonction d'autocomplétion sur le smartphone : ils proposent le mot suivant. Mais les prouesses et les progrès des modèles de langage les plus modernes tels que GPT-3, LaMDA, PaLM ou ChatGPT sont époustouflants, avec par exemple des programmes informatiques capables d'écrire dans le style d'un poète donné, de simuler des personnes décédées, d'expliquer des blagues, traduire des langues, et même produire et corriger le code informatique – ce qui aurait été impensable il y a quelques mois à peine. Pour faire cela, les modèles se basent sur des modèles de neurones de plus en plus complexes.

Quand les intelligences artificielles parlent à tort et à travers

Ceci dit, les modèles sont plus superficiels que ces exemples nous font croire. Nous avons comparé les histoires générées par des modèles de langage à des histoires écrites par des humains et constaté qu'elles étaient moins cohérentes, mais engageantes, et moins surprenantes que les histoires écrites par les humains.

Plus important encore, nous pouvons montrer que les modèles de langage actuels ont des problèmes même avec des tâches de raisonnement simples. Par exemple, lorsque nous demandons :

« L'avocat a rendu visite au médecin ; le médecin a-t-il rendu visite à l'avocat ? »

… les modèles de langage simples ont tendance à dire oui. GPT3 répond même que l'avocat n'a pas rendu visite au médecin. Une raison possible que nous sommes en train d'explorer est que ces modèles de langage encodent les positions des mots de manière symétrique, et donc ils ne font pas la distinction entre « avant le verbe » et « après le verbe », ce qui complique la distinction du sujet et de l'objet dans une phrase.

De plus, les limites théoriques des modèles de langage basés sur les « transformateurs » signifient qu'ils ne peuvent pas distinguer les séquences paires et impaires d'un certain élément, si celles-ci sont intercalées avec un autre élément. En pratique, cela signifie que les modèles ne peuvent pas résoudre une tâche que nous appelons la « tâche pizza » – une simple énigme de la forme :

« La lumière est éteinte. J'appuie sur l'interrupteur d'éclairage. Je mange une pizza. J'appuie sur l'interrupteur d'éclairage. La lumière est-elle allumée ? »

Ici, une séquence paire d'interrupteurs d'éclairage signifie que la lumière est éteinte, mais un modèle BERT n'arrive pas à l'apprendre. Les modèles les plus puissants actuellement (GPT-3 et chatGPT) refusent catégoriquement de conclure que la lumière est éteinte.

Les modèles de langage d'aujourd'hui ont également des difficultés avec la négation, et réussissent généralement mal les tâches de raisonnement dès que celles-ci sont plus complexes. Par exemple, considérons l'énigme suivante de l'examen national des fonctionnaires de Chine :

« *David connaît l'ami de M. Zhang, Jack, et Jack connaît l'amie de David, Mme Lin. Tous ceux qui connaissent Jack ont une maîtrise, et tous ceux qui connaissent Mme Lin sont de Shanghai. Qui est de Shanghai et a une maîtrise ?* »

Les modèles actuels ne répondent correctement que dans 45 % des cas et chatGPT refuse de répondre… alors que la meilleure performance humaine est de 96 %.

Le modèle ne sait pas qu'il ne sait pas

Très probablement, le principal inconvénient des modèles de langage actuels est qu'ils sont faits pour faire des analogies, mais pas pour mémoriser exactement un ensemble trop grand. Par exemple, ils ne peuvent pas se souvenir d'une liste d'un million de clients ou d'un ensemble de dix mille protéines. Le modèle, à sa propre discrétion, oubliera certains des articles. Un problème encore plus important est qu'il peut aussi, encore une fois à sa propre discrétion, inventer de nouveaux éléments (un processus appelé à juste titre « hallucination »). Cela est une conséquence fondamentale des techniques d'apprentissage automatique des modèles de langage, optimisées pour interpoler et généraliser, pas pour mémoriser.

Le problème est qu'il est impossible de dire à partir de la sortie d'un modèle de langage si une réponse a été mémorisée ou inventée. Par exemple, quand on demande à GPT-3 où est mort Alan Turing, le modèle répond impassiblement « Warrington », au lieu de la réponse correcte (Wilmslow). ChatGPT répond correctement à cette question, mais si on l'interroge sur des chercheurs moins connus, comme les auteurs de cet article, il invente des faits. Et quand on lui pose des questions plus compliquées, il fournit des réponses élégantes, mais fausses :

« L'addition de deux nombres impairs est-elle impaire ? »

… le modèle répond poliment :

« Oui, lorsque vous ajoutez deux nombres impairs, le résultat est toujours un nombre impair. Cela est dû au fait qu'un nombre impair peut s'exprimer sous la forme 2n+1, où n est un entier. Lorsque vous ajoutez deux nombres sous cette forme, le résultat est (2n+1)+(2m+1) = 2(n+m)+2, qui est également sous la forme 2p+1, où p est un entier. Cela signifie que le résultat est un nombre impair. »

Sur la forme, la réponse est très convaincante, mais fausse.

Avec ces exemples, nous souhaitons montrer qu'il reste à l'heure actuelle imprudent de s'appuyer sur un modèle de langage pour raisonner ou prendre des décisions. Les modèles s'améliorent avec le temps, connaissent plus de choses, et savent de plus en plus s'abstenir de répondre lorsqu'ils n'ont pas l'information. Cependant, en dehors des questions simples, un modèle de langage peut facilement inventer une réponse et avec une explication ou une preuve tout aussi inventée et approximative.

D'autres méthodes excellent à raisonner sur des faits exacts

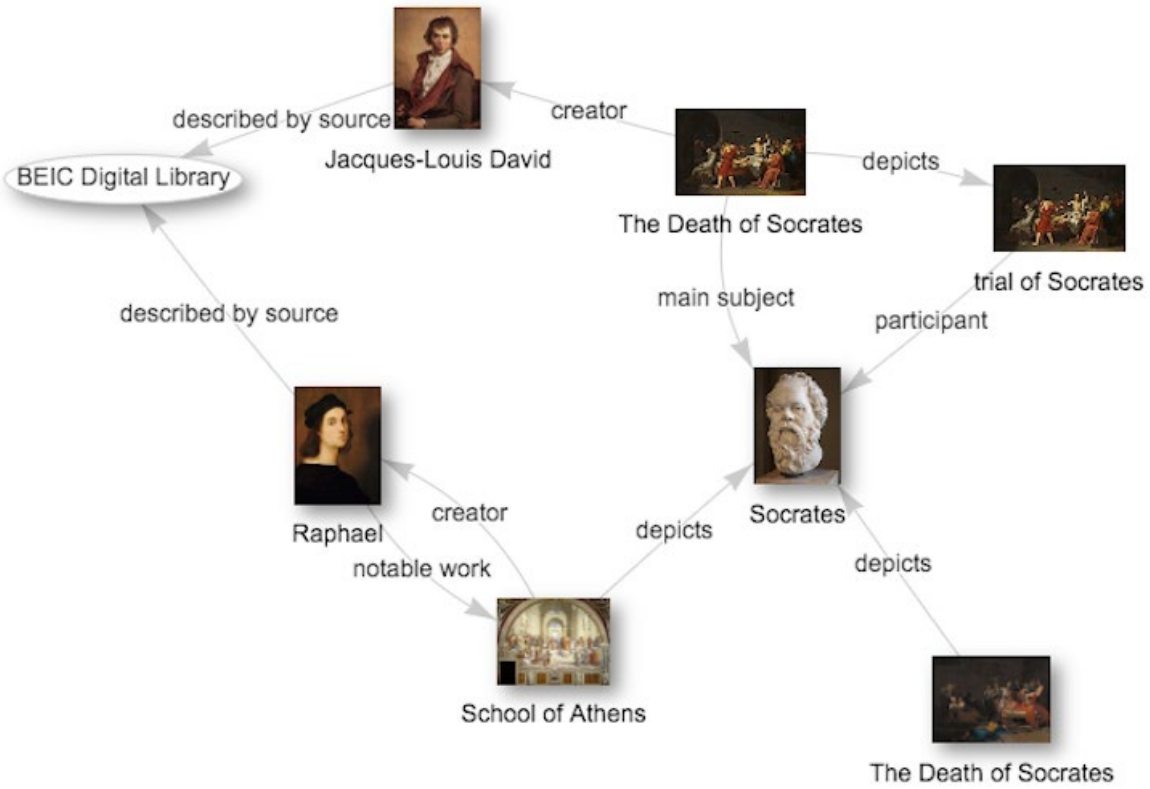
Tout cela ne veut pas dire que les modèles de langage ne seraient pas des outils étonnants aux capacités époustouflantes. Cela ne veut pas non plus dire que les modèles de langage ne pourront jamais surmonter ces défis, ou que d'autres méthodes de *deep learning* ne seront pas

développées à cette fin. C'est plutôt dire qu'au moment d'écrire ces lignes, en 2022, les modèles de langage ne sont pas l'outil de choix pour raisonner ou pour stocker des données exactes.

Pour ces fonctions, l'outil de prédilection reste actuellement les « représentations symboliques » : les bases de données, les bases de connaissances et la logique. Ces représentations stockent les données non pas de façon implicite, mais comme des ensembles d'entités (telles que des personnes, des produits commerciaux ou des protéines) et des relations entre ces entités (telles que qui a acheté quoi, ce qui contient quoi, etc.). Des règles logiques ou des contraintes sont ensuite utilisées pour raisonner sur ces relations d'une manière prouvée correcte – bien que généralement sans tenir compte des informations probabilistes. Un tel raisonnement a par exemple été utilisé dès 2011 par l'ordinateur Watson, lors du jeu Jeopardy pour répondre à la question suivante :

« Quel est le roi espagnol dont un portrait, peint par Titien, a été volé avec arme d'un musée argentin en 1987 ? »

En effet, la question peut se traduire par des règles de logique applicables sur une base de connaissance, et uniquement le roi Philip II peut correspondre. Les modèles de langages ne savent actuellement répondre à cette question, probablement parce qu'ils n'arrivent pas à mémoriser et manipuler suffisamment de connaissance (liens entre des entités connues).



Un exemple très simple de « graphe de connaissance ». Ces objets permettent de connecter des concepts et des entités. Ils sont beaucoup utilisés par les moteurs de recherche et les réseaux sociaux. Fuzheado/Wikidata, CC BY-SA

Ce n'est sans doute pas un hasard si les mêmes grandes entreprises qui construisent certains des modèles de langage les plus puissants (Google, Facebook, IBM) construisent également certaines des plus grandes bases de connaissances. Ces représentations symboliques sont aujourd'hui souvent construites par l'extraction d'information d'un texte en langage naturel, c'est-à-dire un algorithme essaie de créer une base de connaissances en analysant des articles de presse ou une encyclopédie. Les méthodes qui sont utilisées pour cela sont en l'occurrence les modèles de langage. Dans ce cas, les modèles de langage ne sont pas l'objectif final, mais un moyen de construire les bases de connaissances. Ils sont adaptés pour ça parce qu'ils sont très résistants au bruit, à la fois dans leurs données d'apprentissage et dans leurs entrées. Ils sont donc très bien adaptés pour traiter les entrées ambigues ou bruyantes, omniprésentes dans le langage humain.

Les modèles de langage et les représentations symboliques sont complémentaires : les modèles de langage excellent dans l'analyse et la génération de texte en langage naturel. Les méthodes symboliques sont l'outil de choix lorsqu'il s'agit de stocker des éléments exacts et de raisonner dessus. Une analogie avec le cerveau humain peut être instructive : certaines tâches sont suffisamment faciles pour que le cerveau humain les exécute inconsciemment, intuitivement, en quelques millisecondes (lire des mots simples ou à saisir la somme « 2 + 2 ») ; mais des opérations abstraites nécessitent une réflexion laborieuse, consciente et logique (par exemple mémoriser des numéros de téléphone, résoudre des équations ou déterminer le rapport qualité/prix de deux machines à laver).

Daniel Kahneman a dichotomisé ce spectre en « Système 1 » pour le raisonnement subconscient et en « Système 2 » pour le raisonnement avec effort. Avec la technologie actuelle, il semble que les modèles de langage résolvent les problèmes du « Système 1 ». Les représentations symboliques, en revanche, sont adaptées aux problèmes du « Système 2 ». Au moins pour l'instant, il apparaît donc que les deux approches ont leur raison d'être. Qui plus est, tout un spectre entre les deux reste à explorer. Des chercheurs explorent déjà le couplage entre modèles de langage et bases de données et certains voient l'avenir dans la fusion des modèles neuronaux et symboliques en approches « neurosymboliques ».