La Recherche
Mardi 18 mars 2025

DeepSeek : quand une IA découvre comment raisonner/ décryptage par Nathanaël Fijalkow, spécialiste d'apprentissage machine au CNRS, affilié au Laboratoire bordelais de recherche en informatique (université de Bordeaux - CNRS - INP Bordeaux - Inria)

Début janvier 2025, la start-up chinoise DeepSeek frappe un grand coup dans le monde de l'intelligence artificielle en publiant une série de modèles de langage aussi performants que les leaders du secteur, tout en étant moins gourmands en énergie. Au-delà de l'efficacité, c'est leur méthode d'entraînement innovante qui intrigue : leurs ingénieurs affirment avoir un modèle capable de raisonner… sans jamais lui avoir montré de raisonnement. Comment est-ce possible ? Décryptage par Nathanaël Fijalkow, spécialiste d'apprentissage machine au CNRS, affilié au Laboratoire bordelais de recherche en informatique (université de Bordeaux - CNRS - INP Bordeaux - Inria).

Quand déferle l'application DeepSeek en début d'année, une nouvelle IA générative proposée par la société chinoise du même nom, les réactions sont vives. Si le président américain Donald Trump qualifie cette avancée de «signal d'alarme» pour l'industrie technologique américaine, l'action de l'entreprise Nvidia, principal fabricant de puces pour l'IA chute brusquement de 17 %, entraînant des pertes colossales (près de 600 milliards de dollars en une journée). Pourquoi cette crise? D'abord parce que la série de grands modèles de langue publiés par cette petite entreprise chinoise ont des résultats comparables à ceux qui existent, mais surtout parce que l'architecture de ces modèles les rend moins gourmands en consommation énergétique, et donc moins chers à utiliser.

Pour commencer, il faut bien comprendre l'objectif de DeepSeek: entraîner un modèle capable de "raisonner", c'est-à-dire de suivre des étapes logiques pour résoudre un problème. Les implications philosophiques et techniques sont nombreuses, et dépassent le cadre de cet article. Ici, raisonner est mesuré de manière très terre à terre, par la capacité à résoudre des problèmes de mathématiques et à écrire du code informatique. Il existe en effet de nombreux "benchmarks" (bases de tests de questions et de réponses) permettant d'évaluer la capacité des modèles à résoudre des problèmes issus des olympiades de mathématique et d'informatique par exemple. DeepSeek n'est pas la première entreprise à construire des modèles capables de raisonner. Certains modèles d'OpenAI, par exemple o1, construisent également des raisonnements. De surcroît, toutes les tâches ne nécessitent pas de raisonner. Ainsi pour répondre à la question "Quelle est la capitale de la France?", il suffit de s'appuyer sur ses connaissances. Bien que construire des modèles capables de raisonner ne soit utile que pour certaines tâches, c'est considéré comme un objectif important en IA.

L’objectif de cet article n’est pas de discuter des aspects économiques ni des aspects légaux et éthiques (en particulier, quelles données et modèles ont été utilisés pour l’entraîner, et quelles données personnelles sont récoltées en utilisant l’application de DeepSeek), mais seulement d’expliquer, avec le moins de jargon possible, la contribution scientifique de DeepSeek. Il faut souligner que DeepSeek a fait un choix audacieux: d’une part, elle a publié depuis sa création de nombreux articles détaillant les techniques qu’elle a développées (disponibles en ligne sur la plate forme ArXiv) et, d’autre part, tous les modèles qu’elle a créés sont ouverts, ce qui signifie que chacun peut les télécharger gratuitement, les utiliser, les évaluer, les entraîner… Ce choix de la science ouverte, peu courant parmi les géants de l’IA, est la raison pour laquelle la communauté scientifique a pu comprendre, s’approprier et reproduire les résultats de DeepSeek. Ainsi, l’entreprise franco-américaine Hugging Face a répliqué les expériences, et obtient des conclusions similaires.

Les bons résultats de DeepSeek sont dus à une combinaison d’idées. Certaines concernent l’architecture du modèle, en particulier le “mixture of experts”, en français l'agrégation d’experts, où le modèle monolithique est scindé en de nombreux modèles plus petits et indépendants, appelés des “experts”. Intuitivement, chaque expert se spécialise dans une tâche, et le modèle peut donc s’appuyer sur ces experts pour résoudre un problème. Des idées proches avaient déjà été élaborées, l’un des principaux intérêts étant d’obtenir des modèles moins coûteux en termes de consommation énergétique. Ici nous nous focaliserons sur la contribution la plus novatrice de DeepSeek, qui concerne la méthode d’apprentissage.

Les méthodes d'apprentissage des modèles de langues sont étonnamment similaires à celles employées par les humains. Par exemple, lorsque ma fille de 5 ans voulait apprendre les additions, j'ai essayé plusieurs méthodes. Pour toutes ces méthodes, l’interaction commence de la même manière : je demande à ma fille de faire une addition, par exemple “12 + 19”, et de m’expliquer son raisonnement et la réponse à laquelle elle aboutit. Ensuite vient une phase d’évaluation et de récompense :

Approche 0 : je récompense la proximité du raisonnement de ma fille avec le mien: “12 + 19 = 31 : j’ajoute d’abord les unités, j’ai une retenue, que j’ajoute aux dizaines…”. Dit autrement, j’apprends à ma fille à faire des additions comme je le fais moi, en lui apprenant à reproduire mon raisonnement.

Approche 1 : j’évalue raisonnement et réponse, et octroie une récompense quand les deux sont corrects. Notez qu’évaluer la réponse est facile (c’est une simple addition!), mais apprécier la qualité d’un raisonnement nécessite bien plus d’efforts, parce qu'il peut être très différent du mien.

Approche 2 : je n’évalue que la réponse: si elle est correcte, je la récompense, sinon je lui signale que la réponse n’est pas exacte.

Retour d’expérience: l’approche 0 est un échec complet, ma fille se désintéresse très vite de mes explications compliquées. L’approche 1 ne fonctionne pas beaucoup mieux: elle n’aime pas trop que j’évalue son raisonnement, encore une fois mes explications sont trop abstraites. L’approche 2 est plus efficace car jour après jour, les résultats s’améliorent! À noter cependant que le raisonnement reste approximatif, même quand les résultats sont corrects puisqu’à aucun moment je ne lui explique comment je raisonne.

L’approche 0 correspond à ce que l’on appelle “supervised fine-tuning”, que l’on pourrait traduire par “entraînement supervisé du modèle”. Ce type d'approche présente deux difficultés: d’abord, il faut obtenir une base de données d’entraînement, ce qui suppose d’avoir plusieurs centaines de milliers de problèmes possédant une solution découlant d’un raisonnement clairement présenté. La seconde concerne l’entraînement du modèle : intuitivement, il est difficile d’apprendre un concept compliqué (l’algorithme d’addition) à partir d’exemples.

L’approche 1 s’appelle “reinforcement learning with human feedback”, que l’on pourrait traduire par “apprentissage par renforcement guidé par un humain”. Dans le cas de ma fille, je suis l'humain qui évalue le raisonnement. Évidemment, ce n'est pas viable à plus grande échelle : pour entraîner des modèles, il faut une très grande quantité de données. D'où l'introduction d'un modèle de récompense (“reward model”): il s’agit de construire un modèle dont l'objectif est d’apprendre à évaluer comme un humain. Une fois ce modèle entraîné, il remplace l’humain dans l’annotation des données. Mais il est très difficile d’entraîner des modèles de récompenses ; on déplace le problème original…

Avant de présenter l’approche 2, expliquons sa motivation: évaluer les raisonnements, que ce soit par un humain ou avec un “reward model”, est coûteux à la fois financièrement et en termes de calcul. La question est donc: comment simplifier l’évaluation des raisonnements ?

L’approche 2 est ce qu’on nomme “reinforcement learning with rule-based feedback”, que l’on pourrait traduire par “apprentissage par renforcement guidé par des règles”. Le principe est simple: puisque le raisonnement est difficile à évaluer, ne l’évaluons pas! On se limite à évaluer seulement la réponse finale. Par exemple, est-elle correcte ou non, indépendamment du raisonnement, etc. C’est un modèle de récompense basé sur des règles. Lorsque c'est possible, on utilise des règles un peu plus avancées, par exemple pour estimer si la réponse est “presque correcte”. C’est infiniment plus simple, mais également bien moins précis qu’un modèle de récompense classique, puisque le raisonnement n’est pas évalué. Mais parfois, cela peut être suffisant !

L’approche de DeepSeek est similaire à l’approche 2, mais elle s’appuie sur un autre aspect des modèles de langues. C’est là que l’analogie entre ma fille de 5 ans et les modèles de langues s’arrête: en effet, ces derniers sont stochastiques, c’est-à-dire que si on lui pose plusieurs fois la même question, on obtient des réponses différentes, chacune illustrant la compréhension de la question par le modèle. À l’inverse, poser plusieurs fois la même question à ma fille aura une conséquence prévisible: un désintérêt rapide et total pour l’exercice.

L’algorithme Group Relative Policy Optimization (GRPO, pour “optimisation de politiques relative à un groupe”) mis au point par DeepSeek dans un article publié mi-2024 s’appuie sur la stochasticité des modèles. On pose 100 fois la même question au modèle et on obtient 100 réponses, que l’on appelle le groupe. On évalue chacune à l’aide de règles (c’est-à-dire l’approche qui n’évalue que la réponse et non le raisonnement), on obtient ainsi pour chaque réponse une récompense sous la forme d’un nombre, disons entre 0 et 1. Ensuite, on calcule la moyenne parmi les cent réponses, et on en déduit l’avantage de chacune des cent réponses. L'avantage est la différence entre la récompense et la moyenne. Une réponse qui a un avantage positif est meilleure que la moyenne, et inversement, s’il est négatif, elle est pire que la moyenne. On entraîne ensuite le modèle de manière classique: intuitivement, on modifie les

paramètres pour augmenter la probabilité de générer les réponses ayant un avantage positif, et réciproquement diminuer la probabilité de générer les réponses ayant un avantage négatif. Plus précisément, dans les deux cas, la probabilité sera modifiée proportionnellement à l'avantage : plus l'avantage est grand, plus la probabilité sera augmentée.

À ce stade, nous pouvons énoncer le résultat extrêmement surprenant de DeepSeek: le modèle R1-zero de DeepSeek, entraîné exclusivement avec cet algorithme, se met à exhiber des capacités de raisonnement au cours de l'entraînement. C'est inattendu! Le modèle n'a jamais vu de raisonnement, puisque l'entraînement consiste seulement à évaluer les réponses, sans le raisonnement. Par exemple, le modèle se met à réévaluer des étapes du raisonnement effectué: il se remet en cause. Les résultats de R1-zéro sur des benchmarks reconnus et réputés difficiles sont comparables à ceux des modèles de pointe. Au cours de l'entraînement, on observe que les raisonnements s'allongent: le modèle comprend l'intérêt de raisonner avant de répondre à la question. On note également que bien que les résultats s'améliorent, le raisonnement reste difficile à comprendre. En fait, le modèle développe un langage interne qu'il utilise pour raisonner, mais puisque personne ne lui a montré comment un humain raisonne, il le fait à sa manière! Notez la similarité avec les observations sur ma fille et les additions: son raisonnement, bien que différent du mien et difficile à comprendre pour moi, est très important pour arriver à la solution, souvent correcte.

DeepSeek va plus loin en construisant un deuxième modèle, R1, qui est entraîné à partir de R1-zero en combinant l'algorithme GRPO avec des approches plus classiques d'entraînement supervisé du modèle ("supervised fine-tuning", l'approche 0). De cette manière, R1 apprend également à raisonner à partir d'exemples de raisonnements, et donc son raisonnement est plus proche de celui d'un humain, et devient plus intelligible. Le modèle R1 est celui qui est commercialisé par DeepSeek à travers son application.

Quel est l'intérêt de l'algorithme GRPO par rapport aux approches précédentes? Sa première force est sa simplicité: il est bien plus facile à implémenter et à mettre en œuvre puisqu'il ne nécessite pas de construire une base de données d'entraînement incluant des raisonnements humains. Pour cette raison, il va probablement être largement adopté pour l'entraînement d'autres modèles. Mais son intérêt principal réside dans cette observation surprenante : il permet d'apprendre à un modèle à raisonner sans jamais montrer d'exemples de raisonnements !

L'arrivée de DeepSeek et de ses modèles ouverts marque un tournant dans la recherche en IA. En montrant qu'un modèle peut apprendre à raisonner sans supervision explicite, la start-up bouscule les certitudes et remet en question les approches dominantes de l'IA. Son pari sur la science ouverte contraste avec la stratégie plus opaque des géants du secteur et pourrait bien redistribuer les cartes. Reste à savoir si cette avancée se traduira par une adoption massive de son algorithme, ou si les mastodontes de l'IA parviendront à absorber et dépasser cette innovation. Quoi qu'il en soit, DeepSeek a déjà réussi à imposer une nouvelle question centrale dans le domaine : et si l'intelligence artificielle apprenait à raisonner… autrement que nous ?

RETOUR AU SOMMAIRE