Les IA vont-elles consumer toute l'énergie du monde ?

Publié le 14 juin 2023

Christopher Gannatti, CFA
Global Head of Research

Avec une question aussi vaste, rien ne demeure constant. Quand on considère une nouvelle technologie :

- On ne peut présupposer que les taux d'adoption ou d'utilisation vont demeurer constants : ils peuvent baisser, ils peuvent même monter.
- On ne peut présupposer que la technologie répondant à nos besoins énergétiques va demeurer constante, il pourrait y avoir des progrès en matière d'efficacité ou des changements dans le bouquet énergétique global.
- On ne peut présupposer que l'efficacité de la technologie spécifique adoptée va demeurer constante, on a vu de nombreux exemples de secteurs dans lesquels la version initiale d'un élément de la technologie ou du logiciel est ensuite améliorée et a de meilleures capacités et une utilisation d'énergie plus faible.

Il faut également reconnaître que l'intelligence artificielle (IA) elle-même pourrait suggérer des améliorations en ce qui concerne l'efficacité énergétique d'applications spécifiques, comme le chauffage et la climatisation d'un bâtiment. Ainsi, toute analyse de l'utilisation énergétique et de l'IA doit reconnaître que la seule constante sera le changement.

L'impact environnemental de grands modèles de langage (LLM) spécifiques

Les LLM ont eu la plus grande partie de l'attention dans l'effervescence actuelle autour de l'IA générative. Il est cohérent de réfléchir à la quantité d'émissions de carbone générées par certains de ces systèmes. Le Stanford AI Index Report, publié en 2023, a fourni des données et a noté que des facteurs tels que le nombre de paramètres dans un modèle, l'indicateur d'efficacité énergétique1 d'un centre de données et l'intensité carbone du réseau ont tous leur importance.

Les Graphiques 1a et 1b fournissent plus de détail concernant l'intensité de dioxyde de carbone dans différents LLM :

- De multiples facteurs déterminent le résultat : dans le Graphique 1a, il est tentant d'observer le « Nombre de paramètres » et de prédire qu'un plus grand nombre de paramètres signifie toujours plus d'émissions. On peut voir que Gopher, à 280 milliards de paramètres, n'avait pas l'équivalent d'émissions de dioxyde de carbone le plus élevé x IEE (cette distinction est revenue à GPT-3).

- GPT-3, bien que n'étant pas le plus grand modèle par nombre de paramètres, a été utilisé à l'intensité de carbone en réseau la plus élevée des modèles montrés et avait la consommation d'énergie la plus importante. C'est sans aucun doute la recette pour des émissions élevées, mais il est intéressant de constater que cela nous force à réfléchir non seulement au nombre de paramètres, mais à la localisation du centre de données où l'entraînement a lieu, à l'intensité de réseau des émissions de carbone du centre de données, et à la consommation d'énergie du modèle.
- Le Graphique 1b est conçu pour prendre une activité relativement abstraite (par exemple l'entraînement d'un LLM) et y lier des émissions de carbone résultant d'activités plus concrètes, par exemple prendre un avion de New York à San Francisco. Cela permet de donner une échelle et de mieux comprendre ce que « GPT-3 est associé à l'équivalent de l'émission de 502 tonnes de dioxyde de carbone » signifie.

Graphique 1a : L'impact environnemental de modèles d'apprentissage machine spécifiques, 2022

Source : Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault, « The AI Index 2023 Annual Report », AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, avril 2023.

Les performances passées ne sauraient être un indicateur fiable des résultats futurs et tout investissement peut perdre de la valeur.

Graphique 1b : L'équivalent des émissions de carbone dioxyde (tonnes) par modèle d'apprentissage machine spécifique et exemples de la vie réelle

Source : Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault, « The AI Index 2023 Annual Report », AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, avril 2023.

Les performances passées ne sauraient être un indicateur fiable des résultats futurs et tout investissement peut perdre de la valeur.

Réfléchir à la consommation d'énergie d'un LLM

Les personnes construisant différents LLM ont de nombreux leviers qu'ils peuvent actionner pour influencer différentes caractéristiques, notamment la consommation d'énergie. Des chercheurs de Google ont proposé une famille de modèles de langage nommée GLaM (Generalist Language Model), qui utilise un « mélange d'experts peu activés ». Si une discussion poussée sur la façon dont ce type d'approche fonctionne va au-delà du cadre de cet article, on peut noter que le plus grand des modèles de GLaM compte 1,2 billion de paramètres. En connaissant seulement cette donnée, on peut supposer que ce modèle consommerait plus d'énergie que tout autre modèle examiné dans le Graphique 1a ou 1b2.

En réalité, le modèle GLaM avec 1,2 billion de paramètre consomme seulement un tiers de l'énergie nécessaire pour entraîner GPT-3 et nécessite seulement la moitié des flops de calcul pour les opérations

d'inférence. Une façon simple d'y réfléchir est de penser que, si le modèle total compte 1,2 billion de paramètres, un jeton d'entrée donné dans le modèle GLaM active seulement un maximum de 95 milliards de paramètres, c'est-à-dire que le modèle entier n'est pas actif pour tous les paramètres. GPT-3, lui, activait tous ses 175 milliards de paramètres à chaque jeton d'entrée3. On peut noter que, même si mesurer la performance des modèles d'IA intervient dans de nombreuses dimensions, le modèle GLaM est capable, de nombreuses façons, d'également surperformer GPT-34.

Conclusion

Pour résumer, la conception des modèles est importante, et si les concepteurs de modèles veulent trouver des façons de maintenir la performance en utilisant moins d'énergie, ils ont de nombreuses options pour ce faire.

1 L'indicateur d'efficacité énergétique (IEE) est utile dans son évaluation de l'efficacité énergétique des centres de données de façon normée. IEE = (quantité totale d'énergie utilisée par un centre de données informatique) / (énergie délivrée à un équipement informatique). Un IEE plus élevé signifie que le centre de données est moins efficace.

2 Source : Du et al. « GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. » ARXIV.org. 1 août 2022.

3 Source : Patterson, David; Gonzalez, Joseph; Hölzle, Urs; Le, Quoc Hung ; Liang, Chen; Munguia, Lluis-Miquel; et al. (2022) : The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.19139645.v4

4 Source : Du et al, 1 août 2022.

Blogs associés

+ 4 points essentiels de la conférence sur l'AI d'EM Tech Digital

+ La fin du SaaSacre et l'essor de l'IA générative

+ Mettez la révolution de l'IA à profit avec WisdomTree

Important Risks Related to this Article

Informations importantes

Communications commerciales publiées dans l¼Espace économique européen (« EEE ») Ce document est publié et approuvé par WisdomTree Ireland Limited, une société autorisée et réglementée par la Central Bank of Ireland.

Communications commerciales émises dans des juridictions en dehors de l¼EEE Ce document est publié et approuvé par WisdomTree UK Limited, une société autorisée et réglementée par la Financial Conduct Authority du Royaume-Uni.

WisdomTree Ireland Limited et WisdomTree UK Limited sont toutes les deux désignées comme « WisdomTree » (le cas échéant). Notre Politique sur les conflits d¼intérêts et notre Inventaire sont disponibles sur demande.

Réservé aux clients professionnels uniquement. Les informations figurant dans ce document sont fournies à titre informatif et ne constituent pas une ore de vente, ou une sollicitation d¼ore d¼achat de titres ou d¼actions. Ce document ne doit pas être utilisé comme fondement d¼une décision d¼investissement. La valeur des investissements peut fluctuer et vous êtes susceptible de perte tout ou partie du montant investi. La performance passée ne constitue pas nécessairement une indication des performances futures. Toute décision d¼investissement doit être fondée sur les informations figurant dans le prospectus approprié et sur des conseils indépendants en matière d¼investissement, fiscaux et juridiques.

L¼application des réglementations et lois fiscales peut souvent conduire à des interprétations diérentes. Tous les points de vue ou opinions exprimés dans cette communication représentent les points de vue de WisdomTree et ne doivent pas être interprétés comme des conseils réglementaires, fiscaux ou juridiques. WisdomTree ne donne aucune garantie ou représentation quant à l'exactitude des vues ou opinions exprimées dans cette communication. Toute décision d'investissement doit être fondée sur les informations contenues dans le prospectus approprié et après avoir sollicité des conseils indépendants en matière d'investissement, fiscaux et juridiques. Ce document n¼est pas et ne doit en aucun cas être interprété comme une publicité ou une ore publique d¼actions ou de titres aux États-Unis ou dans toute province ou tout territoire des États-Unis. L¼introduction, la transmission et la distribution (directes ou indirectes) de l¼original ou d¼une copie de ce document sont interdites aux États-Unis.

Bien que WisdomTree s¼eorce d¼assurer l¼exactitude du contenu de ce document, WisdomTree ne peut en garantir l¼exactitude. Les fournisseurs de données tiers sollicités pour obtenir les informations contenues dans le présent document ne donnent aucune garantie ou représentation de quelque sorte en rapport avec ces données. Lorsque WisdomTree exprime ses propres opinions concernant le produit ou l¼activité du marché, ces opinions sont susceptibles de changer. WisdomTree, ses ailiés et leurs dirigeants, directeurs, partenaires ou employés respectifs déclinent toute responsabilité pour toute perte directe ou indirecte découlant de l¼utilisation de ce document ou de son contenu.