GCH2730

Énergie et développement durable dans les systèmes informatiques

Hiver 2023

Les Modèles de Language

Groupe [1]
Maximiliano Falicoff
2013658

Table of Contents

Partie 1 – Place du développement durable en GIGL

Définition

Pour moi, le développement durable est un concept qui vise à assurer la pérennité des activités humaines en harmonie avec l'environnement et la société. Cela signifie que nous devons trouver des moyens de répondre aux besoins actuels sans compromettre la capacité des générations futures à répondre à leurs propres besoins. Le développement durable repose sur trois piliers interdépendants : l'environnement, l'économie et la société. Il s'agit de trouver un équilibre entre ces trois domaines pour garantir un développement équitable.

Le développement durable implique de prendre des décisions éclairées et responsables dans tous les aspects de notre vie, qu'il s'agisse de la consommation de ressources, des choix de carrière, des investissements. Cela implique également de respecter les droits des autres personnes et des autres espèces vivantes, et de considérer les impacts à long terme de nos actions sur l'environnement et la société. En fin de compte, le développement durable est une approche holistique qui vise à assurer un avenir viable et équitable pour tous.

La place du génie informatique et logiciel au sein du développement durable

Le développement durable repose sur trois pilier, l'Environnement, la Société et l'Économie, commençons premièrement par l'aspect environnementale. Les ingénieurs informatiques peuvent contribuer à réduire l'impact environnemental de l'industrie technologique. Par exemple, ils peuvent concevoir des logiciels et des applications économes en énergie que ça soit par des optimisation logicielles ou matérielles, des centres de données plus efficaces, ou même en développant des solutions logicielles pour des problèmes environnementaux tel que des algorithmes de gestion de l'énergie, des systèmes de gestion de l'eau et des déchets, etc.

Sur l'aspect économique, les ingénieurs informatiques peuvent également contribuer à l'économie durable. Ils peuvent concevoir des technologies qui aident à optimiser les chaînes d'approvisionnement, à réduire les coûts de production, à améliorer la qualité des produits et des services. Les logiciels sont à tout niveau de notre société, donc en tant qu'ingénieurs nous avons un impact potentiel sur touts les aspects économiques sur lequel notre société repose.

Sur l'aspect de la société, les ingénieurs informatiques peuvent concevoir des applications qui favorisent la participation civique, le dialogue et la collaboration, l'éducation, la santé, etc. On peut prendre par exemple les réseaux sociaux ou des systèmes de vote à distance. Les technologies de l'information peuvent également aider à résoudre des problèmes sociaux complexes, tels que la pauvreté, l'accès à l'eau potable, l'accès à l'éducation, telles que les systèmes de Massive Open Online Course (MOOC) (Chai & Ivy, s. d.)

Partie 2 – Les Modèles de Langages et leurs impacts sur le développement durable

Les modèles de langage (LLMs) (*Introduction to Large Language Models*, s. d.) sont un type de système d'intelligence artificielle conçu pour comprendre et générer du langage naturel. Ils peuvent être utilisés pour une variété de tâches, telles que la traduction de langues, la synthèse de texte et les chatbots. Les grands modèles de langage sont un développement récent qui a révolutionné le domaine de l'intelligence artificielle.

Commençons par explorer l'histoire des modèles de langage. L'histoire des grands modèles de langage est un domaine relativement récent, datant des années 2010, lorsque les chercheurs ont commencé à expérimenter avec des réseaux neuronaux capables d'apprendre à comprendre et à générer du langage naturel. Une grande avancée a été l'introduction des "embeddings de mots" (*Word Embeddings in NLP*, s. d.) qui ont permis aux réseaux neuronaux de représenter les mots sous forme de vecteurs, facilitant ainsi le traitement du langage naturel grâce à des opérations matricielles extrêmement rapides effectuées par certains composants informatiques (GPU).

En 2017, Google a publié un article intitulé ‘’Attention is all you need’’ (Vaswani et al., 2017) qui a introduit le concept de Transformer. Tout d'abord, parlons du concept d'attention. Le problème avec les réseaux de neurones est qu'ils ont peu de mémoire contextuelle. Ainsi, si l'on commence une phrase, vers la fin, le réseau neuronal oubliera le contexte du début. L'attention permet d'attribuer un certain score aux différentes parties de l'entrée, permettant ainsi d'établir une certaine dépendance entre les différentes parties du texte et d'augmenter la mémoire (Rogel-Salazar, 2022).

Depuis la publication de cet article, des innovations ont été introduites à une vitesse inconnue jusqu'à présent. Un exemple d'une entreprise devenue dominante dans cet espace est OpenAI (*OpenAI*, s. d.) une entreprise qui développe des modèles basés sur cette technologie de Transformer et qui a révolutionné le marché avec des produits tels que ChatGPT .

Impact sur l’économie

Une technologie telle que les LLMs a un potentiel similaire à celui qu'internet avait au début des années 80 pour changer le travail et le système économique. Commençons par parler des impacts sur les entreprises. Dans le domaine du GIGL, le fait que des outils tels que ChatGPT ou CopilotX (*Introducing GitHub Copilot X*, 2023) puissent générer du code fonctionnel signifie qu'il y aura un déplacement de ressources des programmeurs de bas niveau. De plus, ces outils augmenteront largement la productivité, non seulement des programmeurs, mais de tous les travailleurs, grâce aux capacités de rédaction et de résumé de texte. On aura alors une transition de ressources humaines, monétaires et matérielles.

Le fait que ces outils soient entraînés sur une grande quantité d'informations provenant d'internet et puissent effectuer des recherches pour être constamment à jour (*Introducing the New Bing*, 2023) signifie que des plateformes journalistiques, telles que le New York Times, ou des moteurs de recherche, tels que Google, verront potentiellement leur trafic diminuer (Marshall, 2023). Ce sont des conglomérats énormes sur l'échelle économique avec une grande quantité d’employés. De plus, nous devons considérer la notion de droit d'auteur, car si l'outil peut aller sur un site et faire un résumé d'un certain article, l'auteur original ne

sera peut-être pas cité ou reconnu pour son travail. En somme, les lois gouvernant le droit d'auteur devront être mises à jour, et cela deviendra un problème politique qui sera abordé dans l'aspect social.

Le fait que ces modèles soient construits sur une très grande base de données à partir d'internet a limité notre capacité, en tant qu'individus et même en tant qu'entreprises, à réaliser toute la chaîne de développement locale pour un tel outil. Nous devons dépendre de services tels que Amazon Web Services (AWS) pour leur puissance de calcul afin de développer des modèles personnalisés. Étant donné que cette technologie prendra de plus en plus de place dans notre quotidien, avec des individus qui voudront personnaliser leurs modèles sur un sujet particulier ou des entreprises sur leur documentation interne, nous allons observer un virage encore plus marqué vers ces grands centres de calculs et de stockage.

Les coûts élevés en énergie des grands modèles de langage peuvent également constituer un poids financier. Les coûts de fonctionnement et de création de ces modèles peuvent être prohibitifs pour les petites entreprises et les organisations, qui peuvent avoir du mal à rivaliser avec les plus grandes organisations qui peuvent se permettre d'investir dans l'infrastructure informatique nécessaire. Si nous prenons GPT3 comme exemple, on sait que entrainer un modèle équivalent couterait entre 500 000 $ et 250000$ (Patel, 2023) et couterais environ 100000$ par jour pour fonctionner(Illinois, 2023).

Les exigences de stockage des grands modèles de langage peuvent représenter un défi pour les organisations et individus qui doivent stocker et gérer ces modèles. Stocker ces modèles peut nécessiter un investissement important dans l'infrastructure de stockage, y compris du matériel et des logiciels spécialisés.

Pour répondre à ces préoccupations, les chercheurs explorent différentes approches pour réduire les exigences de stockage des grands modèles de langage. Une approche consiste à développer des modèles plus compacts qui peuvent atteindre des performances similaires avec moins de paramètres. Un exemple récent de cela est les modèles LLaMA de Facebook (*Introducing LLaMA,* 2023), qui ont été publiés au public, ainsi que les modèles Alpaca de l'Université Stanford (Taori, 2023), qui ont été entraînés avec 7 milliards de paramètres. Ces modèles peuvent être téléchargés et ajustés localement par les utilisateurs finaux pour leurs propres besoins. Nous ne pouvons que supposer que plus de projets de ce type seront de plus en plus courants à l'avenir, ce qui augmentera utilisation et démocratisation de cette technologie.

Impact sur la société

Inspiré par le film 'Her', dans lequel le protagoniste tombe amoureux d'un assistant personnel d'IA, l'intégration des LLMs dans la scène des rencontres en ligne deviennent une réalité et fait même partie de la réalité dès maintenant avec des services tels que Replika(*Replika Web,* s. d.). L'utilisation de ces agents artificiels dans les relations numériques entraine plusieurs implications. Tout d'abord, étant donné que les LLMs peuvent faciliter une meilleure communication entre les partenaires potentiels en fournissant des incitations à la conversation, en aidant à briser la glace ou même en médiant les désaccords, nous pouvons envisager un scénario dans lequel nous dépendons tellement sur les LLMs pour interagir à notre place que nous commençons à perdre la capacité de le faire nous-mêmes (*Why Artificial Intelligence Could Make Dating Better — And Duller,* 2023).

Deuxièmement, avec la capacité de comprendre et de répondre aux émotions humaines, les LLMs peuvent offrir de la compagnie et du soutien aux personnes qui se sentent seules ou isolées. Nous pouvons envisager un scénario dans lequel les célibataires commencent à devenir émotionnellement dépendants des agents et même à les considérer comme un partenaire romantique et s'isolent encore plus. Au fur et à mesure que les

individus développent des attachements émotionnels aux agents d'IA, des questions se posent quant au consentement, à la manipulation et à la nature véritable de ces relations. La société devra établir des normes et des lignes directrices pour assurer l'utilisation éthique des LLMs dans les rencontres en ligne.

En parlant de la législation et du spectre politique, nous pouvons voir un problème surgir. Les récents progrès ont montré que le domaine évolue si rapidement et que nous découvrons de nouvelles façons d'utiliser cette technologie, par exemple, il y a un article très récent qui crée une sorte de simulation avec 25 agents dans un environnement simulé, et ces agents vivent leur vie à partir d'une simple incitation qui leur donne une identité personnelle, très semblable à la série télévisée Westworld (Park et al., 2023). Cela signifie que la législation doit rattraper ces avancements pour réglementer ce qui peut et ne peut pas être réalisé avec ces technologies, tout comme Internet dans les années 90s. Un problème majeur est également que les personnes qui composent notre corps gouvernemental ne comprennent pas cette technologie ou ses cas d'utilisation et ses impacts donc il est difficile d'orienter la législation dans cette direction. Par exemple au Canada, on a introduit une législation en 2022 introduit juste le début d'une législation future concernant l'IA alors qu'elle est déjà présente et impacte un grand nombre de personnes, ce n'est pas assez rapide comme initiative (*The Artificial Intelligence and Data Act (AIDA) – Companion Document*, 2023)

Au niveau individuel, les LLMs ont un impact en fournissant du contenu et des recommandations personnalisés, en offrant un soutien en matière de santé mentale. Par exemple, il y a déjà eu de l'innovation dans le domaine des soins de santé mentale pour inclure des LLMs dans leur pratique (Kjell et al., 2023). Ils améliorent également la communication grâce aux services de traduction.

En ce qui concerne l'éducation et l'apprentissage, les LLMs peuvent s'adapter au style, aux préférences et au rythme d'apprentissage de chaque élève, garantissant qu'ils reçoivent des ressources et des retours d'information adaptés. Cela peut conduire à de meilleurs résultats d'apprentissage et à une plus grande implication des élèves. De plus, les LLMs offrent un accès facile à la connaissance sur divers sujets, démocratisant l'éducation et la rendant plus accessible aux étudiants du monde entier, quel que soit leur niveau socioéconomique. Cependant on doit noter que ce sont des outils et non et raccourcis pour réaliser le travail nécessaire, il faut envisager des cadriciels scolaires pour sécuriser l'utilisation de ces outils. En plus de soutenir les élèves, les LLMs peuvent améliorer l'expérience d'enseignement en fournissant aux éducateurs des outils avancés pour créer des plans de cours dynamiques et engageants, gérer les activités en classe et évaluer les performances des élèves. Les enseignants peuvent également bénéficier du contenu généré par les LLMs, comme des retours en temps réel sur leurs méthodes d'enseignement ou des suggestions d'amélioration.

LLMs, bien qu'étant très utiles, peuvent avoir certains impacts négatifs. Étant donné que les modèles ont été entraînés sur du contenu Internet, où tout n'est pas nécessairement vérifié, le modèle ne sera pas toujours correct (James, 2023) ou présentent un certain biais dû aux donnes fournis pour créer le modèle. Par conséquent, il y a un potentiel plus important pour que les individus tiennent pour acquis les informations fournies par ces outils, par rapport aux moteurs de recherche traditionnels où nous avons généralement une liste de résultats où les premiers sont généralement corrects, on pourra alors voir littératie numérique diminuer. De plus, les LLMs peuvent imiter la manière de parler d'une personne de telle sorte que nous pouvons combiner différents outils tels que les LLMs avec la génération de voix et les DeepFakes, ce qui peut causer de la confusion, éroder la confiance dans les sources d'information, et même endommager les réputations ou mettre en danger les individus. À mesure que le contenu généré par les LLMs deviennent indiscernable du contenu créé par les humains, les individus peuvent avoir du mal à distinguer le vrai du

faux. Cela pourrait entraîner une baisse des compétences en matière de pensée critique et une plus grande vulnérabilité à la manipulation (Gilmore, 2023).

Impact sur environnement

Les coûts énergétiques de la formation et du déploiement des grands modèles de langage sont une préoccupation majeure. Les modèles requièrent des quantités massives de puissance de calcul et de mémoire, qui consomment des quantités importantes d'énergie. L'impact environnemental de ces coûts énergétiques est une préoccupation importante, surtout à mesure que la taille et la complexité des modèles continuent d'augmenter.

Prenons ChatGPT comme exemple, la compagnie OpenAI ne fournit pas d'estimations de son empreinte carbone. Nous avons deux étapes où nous pouvons évaluer son empreinte carbone. Tout d'abord, nous avons l'étape de formation. Dans l'article suivant (Heikkilä, s. d.), ils estiment qu'en formant un seul modèle à grande échelle (plus de 175 milliards de paramètres), cela équivaut aux émissions totales de cinq voitures américaines, y compris la fabrication. C'est préoccupant car ce domaine est relativement nouveau et nos modèles ne feront que devenir plus grands, et la formation ne se fait pas en une seule fois et nécessite un ou plusieurs tours d'ajustement fin, donc l'estimation est un strict minimum.

Pour la deuxième étapes prenons analysons les requêtes, ChatGPT ne fournit pas ces informations mais nous pouvons les interpoler (Ludvigsen, 2023). Selon l'article, ils comparent l'utilisation de ChatGPT à un autre produit pour lequel nous connaissons la consommation d'énergie, qui s'élève à 0,00396 KWh par requête. Traduisant en français les résultats pour un seul mois d'exécution du modèle : "Pour le mettre en perspective, vous pouvez faire fonctionner une ampoule standard de 7w pendant environ 19 050 ans avec 1 168 200 KWh d'électricité (1 168 200 KWh / (7 / 1000) / 24 / 365)". C'est une quantité extraordinaire d'énergie qui est maintenant utilisée quotidiennement. Mais comme Bloom possède 176 milliards de paramètres (Luccioni et al., 2022), en le comparant à GPT4 qui en possède 175 billions, nous devons supposer que l'empreinte carbone est plutôt une comparaison avec GPT3, qui est encore en usage pour le grand public. On ne peut que s'imaginer de l'empreinte carbone des modèles futurs en se basant sur l'augmentation des paramètres.

Pour répondre à ces préoccupations, les chercheurs explorent différentes approches pour réduire les coûts énergétiques des grands modèles de langage. Une approche consiste à utiliser des matériels plus efficaces, tels que des puces spécialisées conçues pour les applications d'intelligence artificielle. Une autre approche consiste à développer des algorithmes et des méthodes d'entraînement plus écoénergétiques qui peuvent réduire la quantité de calcul requise.

Conclusion

En conclusion, les grands modèles de langage représentent une avancée significative dans le domaine du traitement du langage naturel et ont le potentiel de transformer de nombreuses industries et domaines ainsi que les sociétés, les individus et leurs comportements. Cependant, leurs exigences importantes en matière d'énergie et de stockage posent un défi important en termes de durabilité et d'impact environnemental. Par conséquent, nous devons procéder avec prudence car nous ne connaissons pas l'étendue totale des conséquences que cette technologie pourrait avoir sur l'humanité et sur la planète.

Références

Chai, W., & Ivy, W. (s. d.). *What is a MOOC (massive open online course)?* WhatIs.Com. Consulté 10 avril 2023, à l'adresse https://www.techtarget.com/whatis/definition/massively-open-online-course-MOOC

Gilmore, R. (2023, janvier 4). *From deepfakes to ChatGPT, misinformation thrives with AI advancements : Report - National | Globalnews.ca*. Global News. https://globalnews.ca/news/9386554/artificial-intelligence-democracy-misinformation-eurasia-group/

Heikkilä, M. (s. d.). *We're getting a better idea of AI's true carbon footprint*. MIT Technology Review. Consulté 7 avril 2023, à l'adresse https://www.technologyreview.com/2022/11/14/1063192/were-getting-a-better-idea-of-ais-true-carbon-footprint/

Illinois, U. of. (2023, mars 28). *OpenAI's ChatGPT costs $100k per day to run; accelerators could help*. ScienceBlog.Com. https://scienceblog.com/537121/openais-chatgpt-costs-100k-per-day-to-run-accelerators-could-help/

Introducing GitHub Copilot X. (2023). GitHub. https://github.com/features/preview/copilot-x

Introducing LLaMA : A foundational, 65-billion-parameter language model. (2023). https://ai.facebook.com/blog/large-language-model-llama-meta-ai/

Introducing the new Bing. (2023). https://www.bing.com/new

Introduction to Large Language Models. (s. d.). Cohere AI. Consulté 13 avril 2023, à l'adresse https://docs.cohere.ai/docs/introduction-to-large-language-models

James, V. (2023, février 8). *Google's AI chatbot Bard makes factual error in first demo—The Verge*. https://www.theverge.com/2023/2/8/23590864/google-ai-chatbot-bard-mistake-error-exoplanet-demo

Kjell, O., Kjell, K., & Schwartz, H. A. (2023). *AI-based Large Language Models are Ready to Transform Psychological Health Assessment*. PsyArXiv. https://doi.org/10.31234/osf.io/yfd8g

Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2022). *Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model* (arXiv:2211.02001). arXiv. http://arxiv.org/abs/2211.02001

Ludvigsen, K. G. A. (2023, mars 5). *ChatGPT's Electricity Consumption*. Medium. https://towardsdatascience.com/chatgpts-electricity-consumption-7873483feac4

Marshall, A. (2023, février 11). News Publishers Are Wary of the Bing Chatbot's Media Diet. *Wired*. https://www.wired.com/story/news-publishers-are-wary-of-the-microsoft-bing-chatbots-media-diet/

OpenAI. (s. d.). Consulté 12 avril 2023, à l'adresse https://openai.com/

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative Agents : Interactive Simulacra of Human Behavior* (arXiv:2304.03442). arXiv. https://doi.org/10.48550/arXiv.2304.03442

Patel, D. (2023, janvier 24). *The AI Brick Wall – A Practical Limit For Scaling Dense Transformer Models, and How GPT 4 Will Break Past It*. https://www.semianalysis.com/p/the-ai-brick-wall-a-practical-limit

Replika Web. (s. d.). Replika.Com. Consulté 8 avril 2023, à l'adresse https://my.replika.com

Rogel-Salazar, D. J. (2022, mai 25). *Transformers Models in Machine Learning : Self-Attention to the Rescue*. https://www.dominodatalab.com/blog/transformers-self-attention-to-the-rescue

Taori, R. (2023). *Alpaca : A Strong, Replicable Instruction-Following Model*. https://crfm.stanford.edu/2023/03/13/alpaca.html

The Artificial Intelligence and Data Act (AIDA) – Companion document. (2023, mars 13). Innovation, Science and Economic Development Canada. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Why Artificial Intelligence Could Make Dating Better—And Duller. (2023, février 14). Inverse. https://www.inverse.com/science/future-of-love-ai-matchmaker

Word embeddings in NLP : A Complete Guide. (s. d.). Consulté 11 avril 2023, à l'adresse https://www.turing.com/kb/guide-on-word-embeddings-in-nlp