VERS L'ÉVALUATION ET L'ATTÉNUATION DE L'IMPACT ENVIRONNEMENTAL DES GRANDS MODÈLES DE LANGUES

DR. SASHA LUCCIONI

SEPTEMBRE 2023

CIFAR

Réflexions sur l'IA

À PROPOS DES AUTEUR

Crédit photo : Joseph Viviano

DR. SASHA LUCCIONI

Responsable du climat, Hugging Face
sasha.luccioni@huggingface.co

RECONNAISSANCE DES TERRES

Nous reconnaissons les droits territoriaux des Autochtones sur les terres où le CIFAR est présent. Pendant des milliers d'années, ces terres ont été le territoire ancestral d'un grand nombre de peuples tels que les Mississaugas de la Credit, les Anishnabeg, les Chippewa, les Haudenosaunee et les Wendat, et abritent aujourd'hui diverses Premières Nations et populations inuit et métis. Nous sommes reconnaissants de pouvoir y établir le siège de nos activités. Nous sommes aussi conscients que la réconciliation est l'affaire de tous et de toutes. Le programme IA et société du CIFAR vise à enrichir notre compréhension des retombées sociétales de l'IA dans le but de poser les jalons d'une IA responsable. Et l'avenir de l'IA responsable passe par la prise en compte des préoccupations des communautés autochtones. Le CIFAR s'engage à privilégier le point de vue des peuples autochtones dans le développement et la conception d'une IA responsable.

CLAUSE DE NON-RESPONSABILITÉ

Cet exposé n'exprime que les opinions de l'auteure; il s'appuie sur ses recherches antérieures ainsi que sur des recherches provenant d'autres sources, qui sont citées dans les références.

TABLE DES MATIÈRES

2 RÉSUMÉ

3 INTRODUCTION

6 CONSTATATIONS

8 ÉTUDE DE CAS : LE MODÈLE BLOOM

12 DISCUSSION POLITIQUE

14 CONCLUSION

RÉSUMÉ

Ces dernières années, la taille des grands modèles de langues (GML) a augmenté de manière exponentielle. Ainsi, les modèles les plus récents comme GPT-3[1] et OPT[2] sont entraînés à partir de centaines de milliards de paramètres. L'entraînement de modèles de cette taille nécessite généralement des millions d'heures de calcul, un processus qui consomme de grandes quantités d'énergie et émet de nombreuses tonnes de dioxyde de carbone. Malgré cela, peu de publications sur les GML diffusent publiquement le coût environnemental de l'entraînement des modèles ou les journaux d'exploitation nécessaires pour réaliser les évaluations du coût environnemental. En outre, d'autres processus à forte intensité carbonique faisant partie du cycle de vie du modèle sont sous-explorés, tels que la fabrication du matériel spécialisé nécessaire à l'entraînement des modèles, ainsi que les exigences informatiques (et les émissions qui en découlent) de leur déploiement et de leur maintenance à grande échelle. Dorénavant, la mise en place de normes relatives à l'impact environnemental des GML et une plus grande transparence de la part des créateurs de modèles sont des mesures essentielles pour que les utilisateurs de modèles puissent prendre des décisions plus éclairées.

1.0

INTRODUCTION

L'intelligence artificielle (IA) a été l'une des plus grandes avancées technologiques des dernières années. Les avancées fondamentales réalisées dans ce domaine par les entreprises et le milieu universitaire sont utilisées dans une multitude d'applications et d'outils au service des consommateurs, qu'il s'agisse d'outils de recherche sur le Web ou d'appareils intelligents. Les modèles d'IA les plus récents, tels que GPT-4, ont acquis une popularité considérable grâce à leurs capacités génératives; ils ont été utilisés des dizaines de millions de fois dans les premiers mois suivant leur lancement.

Afin de mieux comprendre la durabilité de ces technologies, il est utile de comprendre comment ces modèles sont créés et déployés. Dans cet exposé de politique, nous commencerons par une brève description de l'évolution des modèles d'IA au cours des dernières décennies et du changement de paradigme qui s'est produit avec l'avènement des modèles génératifs ces dernières années. Nous aborderons les différentes étapes du cycle de vie des modèles génératifs, depuis la fabrication des processeurs graphiques utilisés pour leur entraînement jusqu'à leur déploiement dans les produits et services. Nous présenterons ensuite des résultats récents concernant l'empreinte carbone de différents modèles d'IA ainsi que des processus de pré-entraînement et de réglage fin. Nous terminerons avec des recommandations sur les mesures à prendre pour assurer la durabilité des modèles d'IA en recherche et développement.

[1] Brown, T. et al., (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[2] Zhang, S. et al., (2022) OPT: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068

1.1

L'ÉVOLUTION DES GRANDS MODÈLES DE LANGUES

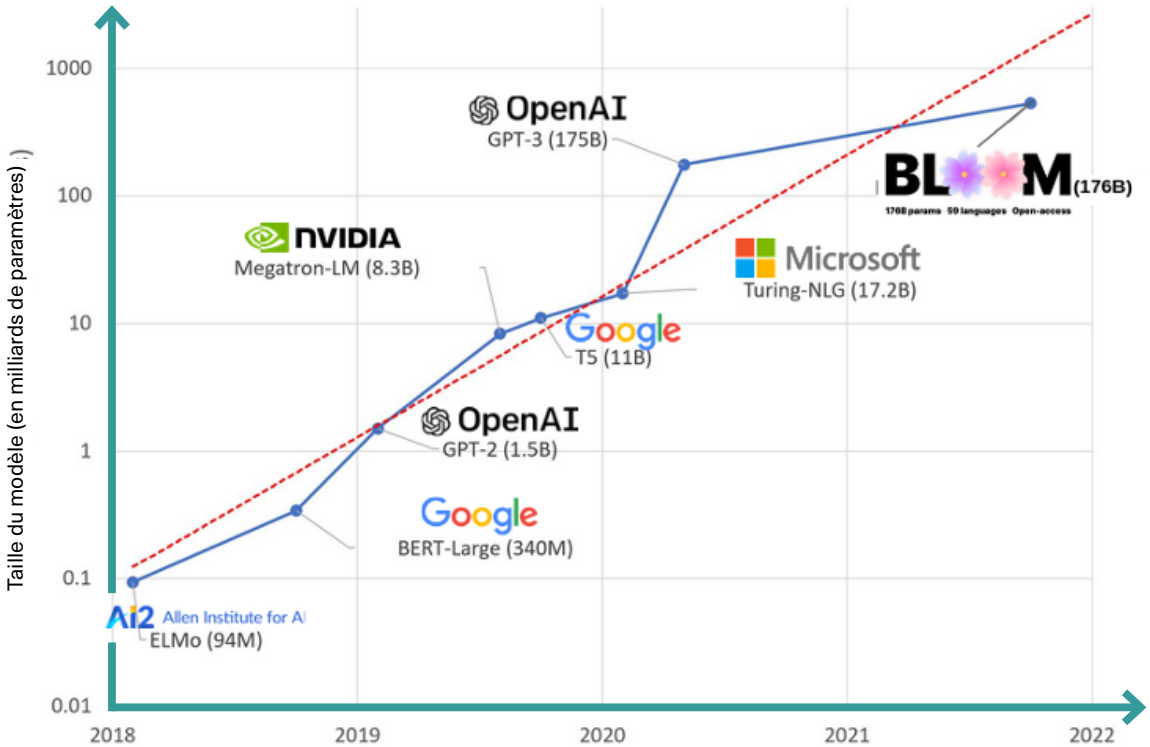
Pour commencer, il est important d'expliquer ce que sont les grands modèles de langues (GML), comment ils fonctionnent et comment ils sont entraînés. Les GML (également appelés modèles fondateurs[3]) sont des réseaux neuronaux autoattentifs[4] qui sont entraînés à partir de grandes quantités de textes, ce qui leur permet de construire des modèles statistiques de la langue. En pratique, cela signifie qu'avec une quantité suffisamment importante de données texte générales (par exemple, l'ensemble de Wikipédia), les GML apprendront à prédire le mot le plus probable à partir d'une saisie de texte. Par exemple, pour « Le chat s'est assis sur _____ », le GML indiquera « la chaise » ou « le tapis ». C'est ce qu'on appelle l'étape de pré-entraînement, qui peut être suivie d'une étape de réglage fin (par exemple, pour adapter un GML à un domaine particulier en l'entraînant à partir des données de ce domaine) ou d'une étape d'instruction (par exemple, pour entraîner un GML à mieux répondre aux commentaires et aux instructions de l'humain). Le GML qui en résulte peut ensuite être déployé dans un outil ou un système d'IA, par exemple un outil tel que ChatGPT, ou être interrogé directement par le biais d'une interface utilisateur ou d'un programme.

Ces dernières années ont été marquées par une augmentation constante de la taille des GML. Celle-ci est souvent déterminée par le nombre de paramètres (c'est-à-dire de connexions) répartis sur les milliers de couches qui constituent le modèle. Il y a quelques années, ces paramètres se mesuraient en millions, comme BERT-Large, un GML publié par Google Research en 2018 et comptant 340 millions de paramètres. Aujourd'hui, BLOOM, le premier GML en libre accès entraîné en collaboration par le projet Big Science et diffusé en 2022, compte 176 milliards de paramètres, soit plus de 5 000 fois plus qu'il y a 4 ans.

[3] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[4] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

FIGURE 1 Évolution de la taille des grands modèles de langues de 2018 à 2022.



L'augmentation de la taille des GML n'est pas un phénomène isolé; elle s'est aussi accompagnée d'améliorations matérielles. Chaque nouvelle génération de processeurs graphiques, ou GPU (pour Graphics Processing Units), c'est-à-dire des puces utilisées pour l'entraînement des GML, est plus puissante que la précédente, non seulement en ce qui a trait à la vitesse, mais aussi à la quantité de connexions entre les différentes couches du GPU. Ce dernier point est particulièrement important pour les GML, car ils impliquent des milliards de calculs à chaque étape de pré-entraînement, et la parallélisation permet de mettre à jour la représentation interne du modèle sur plusieurs couches à la fois. Les générations les plus récentes de GML sont entraînées pendant des milliers d'heures sur des milliers de GPU, ce qui a amené de nombreux scientifiques à s'inquiéter de la viabilité de telles exigences de calcul à long terme[5].

La croissance et le développement des GML ont aussi été marqués par l'accès accru à des collections de données suffisamment importantes pour permettre leur entraînement. Par exemple, alors que le modèle BERT-Large a été entraîné en 2018 à partir d'environ 16 Go de données composées de livres et de Wikipédia, les nouvelles générations de modèles ont été entraînées en grande partie à partir de données extraites du Web, comme celles de Common Crawl, qui représentent 400 To au total. Ces données sont complétées par des données recueillies à partir de commentaires humains, et tout cela contribue à la qualité des générations de modèles. Outre la taille des modèles et les améliorations matérielles, les données ont été un catalyseur important des progrès récents dans le domaine des GML. Mais ces progrès ont aussi un prix en matière d'impacts environnementaux, que nous décrivons plus en détail ci-dessous.

[5] Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. arXiv preprint arXiv:2007.05558.

2.0

CONSTATATIONS

CALCUL DES ÉMISSIONS DE CARBONE ISSUES DE L'ENTRAÎNEMENT DES GML

La plupart des études d'empreinte carbone réalisées à ce jour dans le domaine de l'IA se sont concentrées sur les émissions liées à la production de l'énergie nécessaire à l'entraînement des modèles. Il s'agit de l'estimation la plus simple, car l'entraînement des modèles a un début et une fin bien définis, même s'il peut prendre des semaines, voire des mois. Il n'existe pas de méthodologie unique et reconnue dans ce domaine, et les estimations varient en fonction des facteurs pris en compte par les créateurs de modèles. Les facteurs pris en compte sont généralement les suivants :

1. la durée d'apprentissage du modèle mesurée en heures;
2. la puissance consommée par le matériel utilisé pour entraîner le modèle;
3. l'intensité carbonique du type d'énergie utilisée pour l'entraînement.

La multiplication de ces trois facteurs fournit une première estimation de ce que l'on appelle la « consommation d'énergie dynamique », puisqu'elle ne tient compte que de l'énergie consommée par le matériel qui entraîne le modèle, et non pas des coûts généraux en matière énergétique (par exemple, le refroidissement du centre de données, le stockage, le réseau). Cette mesure se reflète parfois dans l'indicateur d'efficacité énergétique des centres de données, un facteur moyen qui varie en fonction de l'efficacité du serveur où l'entraînement a eu lieu. Cet indicateur est pris en compte dans certaines estimations de l'empreinte carbone des GML (par exemple, T5[6]), mais pas dans d'autres (par exemple, OPT). OPT and BLOOM, in terms of their size, power consumption and CO_{2eq} emissions.

Le premier article à effectuer ce type d'estimation (Strubell et coll.[7]) a évalué que l'entraînement de BERT, le premier grand modèle de langues, a émis environ 284 tonnes d'équivalent dioxyde de carbone. Les différentes sources d'émissions de carbone sont converties en une seule unité de mesure : les équivalents dioxyde de carbone (éq. CO_2). Ces équivalents sont calculés en comparant le potentiel de réchauffement planétaire (PRP) des différents gaz à effet de serre à celui du dioxyde de carbone (CO_2). Par exemple, le méthane a un PRP 25 fois supérieur à celui du CO_2, ce qui signifie que 1 gramme de méthane équivaut à 25 grammes de CO_2. Ces dernières années, un certain nombre d'études supplémentaires ont été menées sur la consommation d'énergie et les émissions de carbone des GML les plus populaires; nous en présentons les résultats dans le tableau ci-dessous. Nous y comparons la taille, la consommation d'énergie et les émissions d'équivalent CO_2 de quatre GML récents, soit GPT-3, Gopher, OPT et BLOOM.

TABLEAU 1

NOM DU MODÈLE	ANNÉE	TAILLE DU MODÈLE (NB DE PARAMÈTRES)	CONSOMMATION D'ÉNERGIE POUR L'ENTRAÎNEMENT	ÉMISSIONS ÉQ. CO_2 PENDANT L'ENTRAÎNEMENT
GPT-3	2020	175B	1,287 MWh	502 tonnes
Gopher	2021	280B	1,006 MWh	352 tonnes
OPT	2022	175B	324 MWh	70 tonnes
BLOOM	2022	176B	433 MWh	25 tonnes

Comparaison des grands modèles de langues des dernières années en fonction de leur taille, de leur consommation d'énergie et des émissions de carbone pendant l'entraînement.

Nous pouvons constater que si la taille des modèles est restée sensiblement la même au cours des deux dernières années (Gopher étant légèrement plus gros que ses contemporains), leur consommation d'énergie est en baisse : l'entraînement d'OPT a nécessité environ 75 % moins d'énergie que celui de GPT-3. Cela est dû en grande partie aux nouvelles générations de matériel informatique, plus efficaces et moins gourmandes en énergie (comme le notent Patterson et coll. en 2022[8]). Il en va de même pour les émissions de carbone des modèles, qui ont diminué d'un facteur d'au moins 20 entre GPT-3 et BLOOM. Cela peut être attribué au type d'énergie utilisé par les deux modèles, GPT-3 a été largement alimenté par le charbon et le gaz naturel, tandis que BLOOM a utilisé l'énergie nucléaire. Cela illustre le potentiel des énergies renouvelables dans l'amélioration de la durabilité des GML; elles peuvent avoir un impact important sur les émissions globales de carbone pendant l'entraînement des modèles, même si la durée d'entraînement continue à se mesurer en millions d'heures de GPU.

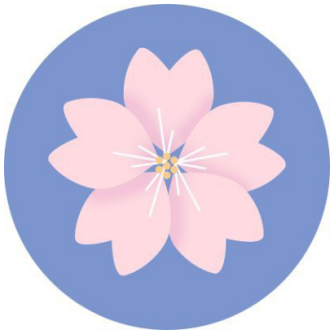
[6]Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

[7]Strubell, E., Ganesh, A., & McCallum, A. (2019, July). Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650).

[8]Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L. M., ... & Dean, J. (2022). The carbon footprint of machine learning training will plateau, then shrink. Computer, 55(7), 18-28

ÉTUDE DE CAS :

LE MODÈLE BLOOM

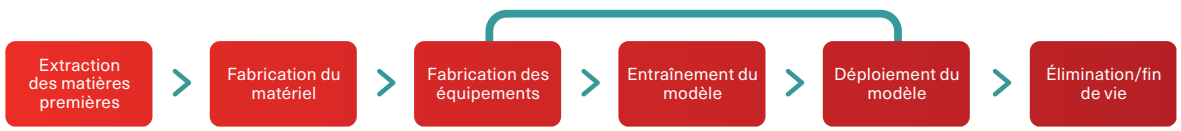


Le modèle de langue multilingue BigScience Large Open-science Open-access (BLOOM) compte 176 milliards de paramètres. Il a été entraîné sur 1,6 téraoctet de données dans 46 langues naturelles et 13 langages de programmation dans le cadre de l'atelier BigScience (en anglais), une initiative d'un an qui a rassemblé plus d'un millier de membres de la communauté de recherche mondiale. Grâce à une généreuse subvention pour frais de calcul du gouvernement français, BLOOM a été entraîné pendant 3 mois sur plus de 1000 GPU, ce qui représente plus d'un million d'heures de calcul GPU.

MÉTHODOLOGIE

S'éloignant un peu de l'entraînement des GML décrit dans la section précédente, il est également intéressant d'effectuer une analyse du cycle de vie (ACV)9 des modèles de langues, qui évalue les ressources requises à tous les stades de leur cycle de vie. Ces ressources comprennent la consommation d'énergie pendant l'entraînement et le déploiement des modèles, mais aussi l'impact sur la planète lié à l'extraction des métaux rares et à la fabrication du plastique requises pour la création du matériel, ainsi que l'eau nécessaire au refroidissement des centres de données qui alimentent les modèles.

FIGURE 2



Les phases prises en compte dans l'évaluation du cycle de vie des GML

Lorsqu'il s'agit d'estimer l'empreinte carbone des modèles d'IA, il manque d'information sur les différentes étapes du cycle de vie des modèles. Par exemple, pour quantifier les émissions engendrées par la fabrication des équipements utilisés pour l'entraînement des GML (GPU, serveurs, etc.), il est nécessaire de collecter les données sur les émissions de portée 3 auprès des concepteurs de ces équipements, ce qui signifie qu'ils doivent à leur tour collecter ces données auprès de leurs fournisseurs. En raison du manque de données concernant les sources d'émissions aux différentes étapes du cycle de vie d'un modèle d'IA, l'estimation de l'empreinte carbone de BLOOM est la première de son genre à aller au-delà de l'entraînement du modèle et à prendre en compte d'autres phases du processus d'analyse du cycle de vie.

[9]Klöpffer, W. (1997). Life cycle assessment: From the beginning to the current state. Environmental Science and Pollution Research, 4, 223-228.

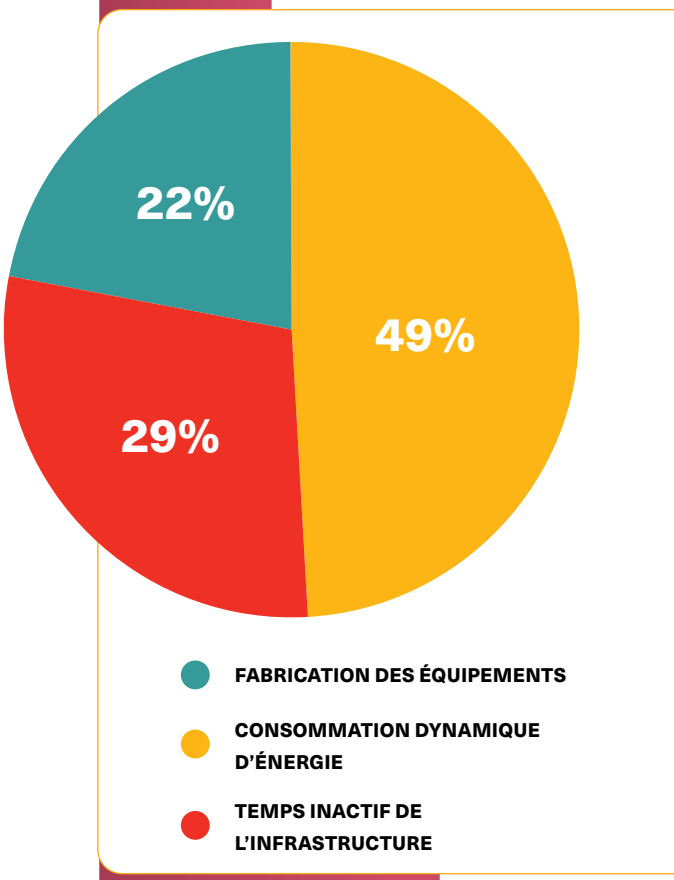
ESTIMATION DES ÉMISSIONS LIÉES À LA FABRICATION, À L'ENTRAÎNEMENT ET À L'INFRASTRUCTURE

D'après nos estimations, la fabrication des équipements a produit 11 t. éq. CO_2, la consommation dynamique pendant l'entraînement du modèle a représenté 25 t. éq. CO_2 et l'infrastructure a produit 14 tonnes supplémentaires pendant les périodes d'inactivité (temps inactif), soit un total de **50 t. éq. CO_2.**

Afin d'estimer les émissions produites par la fabrication du matériel informatique utilisé pour l'entraînement du modèle BLOOM, nous nous sommes basés sur les estimations fournies par les fabricants de matériel tels que HPE, car la quantité exacte pour les GPU Nvidia utilisés pour l'entraînement n'a pas été fournie par le fabricant. Étant donné que l'entraînement de BLOOM a duré au total 1,08 million d'heures en utilisant, en moyenne, 384 GPU sur 48 noeuds de calcul, nous pouvons estimer que la fabrication a produit environ **11,2 t. éq. CO_2.** Cela implique un taux de remplacement tous les 6 ans et une utilisation moyenne de 85 %, ce qui peut varier en fonction de la fréquence des mises à jour du matériel et de l'intensité de son utilisation. Toutefois, compte tenu de l'importante quantité de matériel requise (article en anglais) pour l'entraînement des GML, de nombreuses organisations achètent des dizaines de milliers de GPU récents et se débarrassent des versions plus anciennes et moins efficaces, ce qui augmente leur empreinte carbone globale de façon importante.

En outre, afin de calculer les émissions des différents composants informatiques du centre de données, nous avons mené une série d'expériences, en comparant la consommation d'énergie totale des dispositifs de la grappe de calcul (par exemple, réseau, GPU, stockage, refroidissement/chauffage et noeuds de calcul) lorsqu'ils sont inactifs à la consommation totale des mêmes dispositifs lorsqu'ils sont dynamiques, c'est-à-dire lorsqu'ils exécutent le programme d'entraînement du modèle. Sur cette base, nous avons constaté qu'environ 54 % de la consommation d'énergie peut être attribuée à l'exécution du programme alors que les 46 % restants sont utilisés pour maintenir les noeuds de calcul sous tension. Cela signifie que les estimations présentées ci-dessus, qui se concentrent uniquement sur la consommation dynamique d'énergie, ne tiennent pas compte de près de la moitié de l'énergie, qui est utilisée pour alimenter le reste de l'infrastructure informatique. Ce chiffre dépendra de la génération de matériel et de l'efficacité de l'infrastructure informatique, mais étant donné la taille des GML, ceux-ci sont le plus souvent entraînés de manière distribuée sur des centaines, voire des milliers de noeuds informatiques, ce qui entraîne encore plus de temps inactif. Toutefois, des travaux menés par Google DeepMind (article en anglais) visent à réduire la consommation énergétique pendant le temps inactif en recourant à l'IA, et cette réduction peut atteindre 40 %.

FIGURE 3

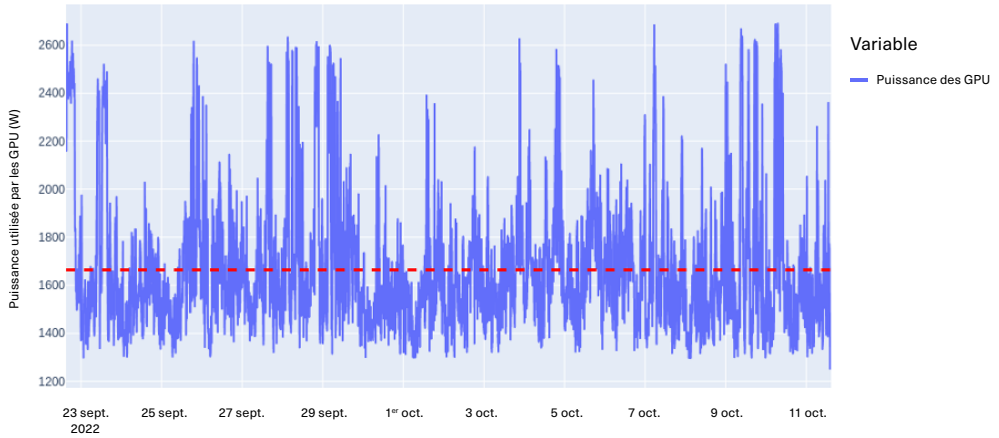


Proportions des émissions de carbone du modèle BLOOM

ESTIMATION DES ÉMISSIONS LIÉES AU DÉPLOIEMENT D'UN MODÈLE

Un article de 2019 (en anglais) estime que de 80 à 90 % de la charge de travail d'apprentissage automatique de Nvidia est associée au déploiement de modèles, mais les chiffres réels de l'utilisation de l'énergie et des émissions de carbone de l'inférence sont largement inconnus en raison de la nature distribuée et dynamique du déploiement. À tout moment, des milliers de modèles peuvent être déployés sur plusieurs instances de calcul en nuage, et leur activité peut varier en fonction de la demande de l'utilisateur. Cela signifie qu'à ce jour, la principale pièce manquante de l'empreinte carbone des GML est le déploiement des modèles, qui n'a fait l'objet d'aucune étude d'empreinte carbone. Afin d'estimer les émissions de carbone liées au déploiement de BLOOM, nous avons déployé le modèle sur une instance de calcul en nuage et suivi la consommation d'énergie de l'instance sur une période d'environ 18 jours. Au cours de cette période, le modèle a reçu en moyenne 558 requêtes par heure, qui ont été traitées en temps réel, soit 230 768 requêtes au total. Cela représente une utilisation réaliste d'un GML en temps réel lors de son déploiement dans des applications comme des agents conversationnels, où ils sont censés répondre à un flux constant et variable de requêtes d'utilisateurs.

FIGURE 4



Fluctuation de la puissance moyenne utilisée pour alimenter les GPU exécutant le modèle BLOOM et consommation moyenne (1 664 W) indiquée par une ligne pointillée rouge

Comme le montre le graphique ci-dessus, au cours de la période de 18 jours pendant laquelle nous avons effectué notre analyse, la puissance utilisée par le modèle BLOOM a fluctué entre 1 252 W et 2 735 W, et la consommation moyenne (1 664 W) est indiquée par une ligne pointillée rouge. Au total, l'instance utilisée pour l'interface de programmation applicative du modèle BLOOM a consommé 914 kWh d'électricité, mais cette consommation a fluctué en fonction de l'utilisation du modèle. Cependant, 75 % de cette énergie a été consacrée au maintien de BLOOM en mémoire, étant donné que les modèles doivent toujours être prêts à répondre aux requêtes des utilisateurs. En outre, étant donné que l'instance en nuage utilisée pour déployer le modèle BLOOM a une intensité carbonique de 394 g éq. CO2 par kWh, cela s'est traduit par l'émission d'environ 19 kg éq. CO2 par jour de déploiement du modèle, soit 340 kg pendant toute la période. Bien que cette quantité ne semble pas énorme comparée aux 50 tonnes émises pendant l'entraînement du modèle, elle prend des proportions imposantes lorsque des GML comme BLOOM sont déployés dans des applications orientées utilisateur comme la recherche et la navigation sur le Web. De telles applications peuvent être interrogées des millions de fois par jour et nécessitent alors de nombreuses instances de GML déployées en parallèle pour répondre aux requêtes.

3.0

DISCUSSION POLITIQUE

Afin de poursuivre une innovation responsable et durable en matière de GML, les éléments suivants doivent être pris en compte :

1 Création de normes et de cadres pour évaluer et rendre compte de l'empreinte carbone des grands modèles de langues. Comme nous pouvons le voir dans les constatations, il n'existe pas de cadre ou de méthodologie unique pour calculer les émissions de carbone des GML, ce qui rend difficile une comparaison valable des émissions de différents modèles. S'il existe des initiatives telles que la Green Software Foundation, qui visent à élaborer des normes pour réduire l'impact environnemental de l'informatique, il n'y a pas de normes précises pour l'IA en général et les GML en particulier, qui tiennent compte des particularités de ces technologies.

2 Développement d'outils pour estimer avec précision la consommation d'énergie. Bien qu'il existe des outils tels que Code Carbon et Carbon Tracker, la précision avec laquelle ils mesurent la consommation d'énergie réelle est remise en question : des travaux récents[10] ont montré que les mesures effectuées varient en fonction du type de matériel utilisé et ne reflètent pas la véritable consommation d'énergie des appareils (lorsqu'elle est mesurée à l'aide d'appareils physiques). En outre, aucun des outils existants ne peut mesurer correctement la consommation d'énergie du déploiement d'un modèle, car l'intervalle de temps utilisé pour mesurer la consommation n'est pas suffisamment granulaire. Il est important d'améliorer et de calibrer les outils existants et de les harmoniser pour qu'ils correspondent à des normes telles que celles décrites ci-dessus; cela permettra de refléter avec plus de précision la consommation d'énergie et les émissions de carbone de tous les modèles, et ce, à toutes les étapes de leur cycle de vie.

3 Rapports sur l'impact environnemental lors de la diffusion de systèmes basés sur les GML. Le modèle BLOOM est le dernier GML pour lequel nous disposons de renseignements suffisants pour estimer son empreinte carbone de manière significative. Les générations suivantes de GML, telles que LLaMa[11], GPT-4[12] et PaLM 2[13], ne fournissent pas suffisamment de détails sur l'entraînement ou l'architecture du modèle pour estimer leurs émissions de carbone. Cela reflète une tendance, parmi les dernières générations de GML, à accorder moins de permission en fait d'accès et de renseignements que les précédentes (Solaiman, 2022[14]). Le fait d'obliger les créateurs de modèles à fournir suffisamment de renseignements sur l'impact environnemental de leurs modèles, par exemple au moyen d'outils tels que les cartes de modèles[15], constitue une étape importante vers la transparence et la responsabilité.

4 Création d'une certification et d'une évaluation des modèles d'IA. Actuellement, les utilisateurs ne peuvent pas choisir de modèles en fonction de facteurs tels que l'efficacité et la durabilité, étant donné que ces renseignements ne sont pas présentés par les fournisseurs de modèles (voir le point ci-dessus). Toutefois, de nombreuses entreprises et organisations tiennent compte des facteurs ESG (environnementaux, sociaux et de gouvernance d'entreprise) lorsqu'elles prennent des décisions commerciales, et elles déclarent leurs propres émissions dans des rapports ESG. Les émissions des GML constituent des émissions de portée 3 (article en anglais), car elles sont le résultat d'activités provenant d'actifs qui ne sont pas détenus ou contrôlés par l'organisation déclarante. Il est donc essentiel que les organisations reçoivent les renseignements qui leur permettront d'estimer la contribution des GML à leur empreinte globale, mais aussi pour tenir compte de ces renseignements lorsqu'elles font un choix parmi différents produits et services fondés sur les GML.

5 Entraînement des modèles à plus faible empreinte carbone. Des recherches récentes[16] ont montré qu'il existe des moyens de réduire l'empreinte carbone de l'entraînement des GML en tenant compte de l'intensité carbonique marginale de l'énergie utilisée pour l'entraînement du modèle. Ce renseignement n'est pas facilement accessible sur la plupart des plateformes commerciales d'informatique en nuage, ce qui signifie que les utilisateurs ne peuvent pas en tenir compte lorsqu'ils choisissent une instance de calcul. Des outils tels que le tableau de bord de l'impact des émissions de Microsoft peuvent contribuer à une plus grande transparence des émissions dans différentes régions et pour différentes ressources, ce qui peut guider l'entraînement et le déploiement des modèles.

6 Développement des ressources en énergies renouvelables. AComme le montre le tableau comparatif des GML plus haut, le type d'énergie utilisé pour l'entraînement des GML peut contribuer à réduire considérablement l'empreinte carbone de l'entraînement (par exemple, de plus de 70 t. éq. CO_2 pour l'entraînement d'OPT à 25 pour l'entraînement de BLOOM, ceteris paribus). Malgré cela, les plus gros serveurs en nuage du monde sont actuellement situés dans des endroits où les réseaux sont principalement alimentés par le pétrole et le gaz naturel[17], ce qui contribue à l'empreinte carbone élevée de l'entraînement d'un grand nombre de GML. En les entraînant dans des régions telles que le Québec, qui est alimenté par de l'énergie hydroélectrique, cela permettrait de réduire leurs émissions. Toutefois, pour que cela soit réalisable, il faut construire davantage de grappes de calcul dans ces régions.

[10]Bannour, N., Ghannay, S., Névéol, A., & Ligozat, A. L. (2021, November). Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing (pp. 11-21).

[11]Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[12]OpenAI (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

[13]Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). PaLM 2 Technical Report arXiv preprint arXiv:2305.10403.

[14]Solaiman, I. (2023, June). The Gradient of Generative AI Release: Methods and Considerations In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 111-122).

[15]Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

[16]Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., ... & Buchanan, W. (2022, June). Measuring the carbon intensity of AI in cloud instances. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1877-1894).

[17]Greenpeace (2020) How Tech Companies are Helping Big Oil Profit from Climate Destruction

4.0

CONCLUSION

L'impact sur l'environnement des nouvelles générations de technologies d'intelligence artificielle telles que les GML est sous-exploré et souvent négligé dans la recherche et la pratique. Des recherches récentes sur l'empreinte carbone de BLOOM, un modèle de langues de 176 milliards de paramètres, ont montré que la fabrication de l'équipement, l'entraînement du modèle et son déploiement sont responsables de quantités non négligeables d'émissions de carbone, d'autres modèles de taille similaire ayant une empreinte encore plus importante. Des recherches supplémentaires s'imposent pour mieux comprendre les émissions des GML à différents stades de leur cycle de vie, y compris la fabrication de l'équipement et le déploiement. Pour ce faire, il est essentiel de faire preuve d'une plus grande transparence en ce qui concerne les émissions réelles des différents modèles et architectures, et de développer des cadres et outils normalisés pour mesurer et certifier les émissions de carbone des modèles. Alors que les GML sont déployés dans un grand nombre d'outils et d'applications destinés aux utilisateurs, il est plus important que jamais de prendre des décisions qui tiennent compte de facteurs tels que la durabilité et l'efficacité.

CIFAR

MaRS Centre, West Tower
661 University Ave., Suite 505
Toronto, ON M5G 1M1 Canada

www.cifar.ca/ai