Synthèse des articles

Liste des articles :

1. Sustainable AI: AI for sustainability and the sustainability of AI
2. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink
3. KNAS: Green Neural Architecture Search
4. Sustainable AI: An integrated model to guide public sector decision-making
5. A Holistic Framework for Designing Carbon Aware Datacenters
6. Sustainable AI: Environmental Implications, Challenges and Opportunities
7. Energy and Policy Considerations for Deep Learning in NLP
8. Quantifying the Carbon Emissions of Machine Learning
9. Green AI

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Problématique | Méthode  (Parler des analyses dans chaque catégories) | Résultat | Contribution  Personnelle | Critique personnelle  (plus argumenter) | Point d’amélioration | Implémentation  Si il y a quelque chose qu’ils n’ont pas pris en compte dans l’implémentation) |
| 1 | Présenter le concept d'IA durable comme un terme générique pour couvrir deux branches ayant des objectifs et des méthodes différents : l'IA pour la durabilité et la durabilité de l'IA | Article review :   * Montre que l’IA durable est très prometteuse * Explique ce que l’IA durable devrait être * Utilise l’article de Strubell pour montrer qu’il y a des réglementations à imposer. | Recommandations :   * Garanties éthiques doivent être mises en place pour protéger les personnes et la planète * Intégrer la durabilité de l’IA dans la politique * Instaurer un "cadre de proportionnalité" pour évaluer si la formation ou le réglage d'un modèle d'IA pour une tâche est proportionnel son empreinte carbone, à son impact environnemental et social général * Utiliser des outils tels que le "Carbon Tracker" pour suivre l'empreinte carbone de la formation d'un certain modèle ou l'empreinte afin d'arrêter la formation du modèle si le coût environnemental prédit est dépassé | * Une étude de différents articles * Des recommandations * Une définition de l’IA durable en deux parties durabilité de l’IA et L’IA pour le développement durable | Propose une bonne façon d’incorporer la politique dans la gestion de l’IA. Cependant le principe d’AI durable aurait pu être un peu plus développé | Développer la définition et l’explication de l’IA durable. Donner des résultats concrets des outils proposés. |  |
| 2 | Réduire la consommation d'énergie et les émissions de carbone des charges de travail ML | Méthodologie des 4 M :   * Model * Machine * Mécanisation * Map | * Une réduction des émissions de 747x sans perte de précision du Transformer * GLaM peut réduire l'empreinte carbone brute d'environ 14 fois avec une précision accrue par rapport au GPT-3, | Proposition de la méthode des 4 M. | Approche très intéressante mais manque de détail et de méthode sur les études de cas | Développer plus la partie sur l’étude de cas et la refaire avec une meilleure méthodologie | Appliquer la règle des 4M sur la classification des images |
| 3 | Peut-on évaluer les architectures sans entrainement ? | * Montre que les gradients dépendent de la conception des réseaux * Démontrer théoriquement que la gram matrix de gradients GM est une mesure cruciale et complète pour évaluer la qualité des architectures * Montre que MGM a de bonnes corrélations avec la perte de formation et la précision de validation * Propose une solution NAS légère basée sur le kernel, appelée KNAS * Compare cette nouvelle approche avec différentes model déjà existants * Vérification de la capacité de généralisation de KNAS | KNAS a de meilleures performances par rapport à RoBERTA-large avec des améliorations de précision de 1,24 et 0,24 sur les ensembles de données MRPC et RTE. | Ont proposé et validé une hypothèse permettant d’obtenir une version verte du NAS pouvant trouver des architectures performantes avec des ordres de grandeur plus rapides que les solutions NAS "train-then-test" |  |  |  |
| 4 | Comment le concept d'IA durable peut-il être défini et opérationnalisé pour guider les décisions du secteur public ? | Literature review :   * Présentation du contexte de la prise de décision du secteur public en matière de gouvernance de l'IA * Présentation de l’approche de recherche * Comparaison des différentes façons dont l’IA durable est conceptualisée et abordée dans les différents articles. * Identification et analyse des conditions limites nécessaire de l’IA en intégrant la littérature sur la durabilité et sur l'IA du secteur public afin d'aboutir à un modèle conceptuel préliminaire d'IA durable pour le secteur public |  | Modèle des 5 « conditions limites » :   * La diversité * La capacité d’apprendre * La capacité d’auto-organisation * Le sens commun * La confiance | Prend surtout en compte l’impact sociétal de l’IA, ne prends pas du tout en compte l’impact environnemental ou économique. Points primordiaux pour l’IA durable. | Intégrer les questions environnementales et économiques à ce modèle |  |
| 5 | Proposer un cadre et une solution 100% 0 carbone 24/7 | * Collecte et évaluation de données des centres de données de Meta * Etablissement et comparaison des différentes solutions proposées * Prise en compte de l’empreinte carbone opérationnelle et intrinsèque | Théoriquement : couverture opérationnelle de 100% | Différents modèles proposés selon les caractéristiques de production d’énergie des régions où se trouve les centres de données :  Energie renouvelable + Batteries + programmation respectueuse | La source d’énergies renouvelables est limitée à l’éolienne et au solaire, pas de recherches poussées sur la façon de stocker l’énergie, seulement une sorte de batterie proposée, pas d’infos sur les installations utilisées dans les différents datacenters | Elargir les données sur lesquelles l’étude est faire avec plus de régions et plus de méthode de stockage. |  |
| 6 | Présenter une analyse de bout en bout de l'empreinte carbone opérationnelle et intrinsèque pour la formation et l'inférence de l'IA. | * Caractériser l'empreinte carbone de l'IA * Analyse de l’empreinte carbone de différents modèles d'apprentissage automatique représentatifs utilisés chez Facebook et de 7 open-source modèles de grande échelle * Optimisation de l’empreinte carbone à partir d’un co-design hardware-software. | Réduction de plus de 800 fois de l'empreinte carbone opérationnelle obtenue grâce à 3 choix de co-design hardware-software judicieux pour un modèle de langage universel basé sur Transformer. | 3 recommandations :   * Platform-level Caching : le pré-calcul et la mise en cache d'intégrations fréquemment utilisées. En utilisant des dispositifs de stockage RAM et Flash comme caches permet un partage entre les applications et les cas d'utilisation. * GPU acceleration : utilisé du hardware IA spécialisé basé sur les GPU * Algorithmic optimization : Réduire la précision de moitié, utiliser des opérateurs personnalisés pour planifier les étapes de codage dans un seul noyau du module Transformer |  |  |  |
| 7 | Quantifier les coûts financiers et environnementaux approximatifs de la formation d'une variété de modèles de réseaux neuronaux ayant connu un succès récent pour le traitement automatique des langues, afin de sensibiliser les chercheurs. | * analyse de l'énergie nécessaire à l'entraînement d'une variété de modèles populaires de NLP (Transformer, ELMo, BERT, GPT-2), ainsi qu'une étude de cas de la somme complète des ressources nécessaires au développement de LISA y compris tous les réglages (tuning) et l'expérimentation. Pour mesurer l'utilisation de l'énergie les auteurs entraînent les modèles utilisant les paramètres par défaut fournis, et échantillonnent la consommation d'énergie du GPU et du CPU pendant l'entraînement. Chaque modèle a été entraîné pendant un jour maximum. Tous les modèles sont formés sur un seul GPU NVIDIA Titan X, à l'exception d'ELMo qui a été formé sur 3 GPU NVIDIA GTX 1080 Ti. Pendant l'entraînement, un échantillonnage de la consommation d'énergie du GPU est effectué de manière répétée. Les auteurs ont ensuite calculé la moyenne de tous les échantillons. Pour échantillonner la consommation d'énergie du CPU, l’interface Running Average Power Limit d'Intel a été utilisée. Le temps total prévu pour que les modèles s'entraînent jusqu'à la fin a été estimé en utilisant les temps d'entraînement et le matériel indiqués dans les articles originaux * Calcul de la puissance totale pt requise à une instance donnée pendant l'entraînement      * Le résultat obtenu est ensuite multiplié par le CO2 moyen produit (en livres par kilowattheure) pour l'énergie consommée aux États-Unis (EPA, 2018). | * Les TPU sont plus rentables que les GPU pour les charges de travail, les modèles émettent des émissions de carbone substantielles ; l'entraînement de BERT sur GPU est à peu près équivalent à un vol transaméricain. So et al. (2019) rapportent que le NAS atteint améliore sa performance de traduction automatique de l'anglais vers l'allemand de seulement 0,1 BLEU mais avec un coût d'au moins 150 000 $ en temps de calcul à la demande et des émissions de carbone non négligeables. * L'entraînement d'un seul modèle est relativement peu coûteux, cependant le coût de la mise au point (tuning) d'un modèle pour un nouvel ensemble de données, estimé par les auteurs ici à 24 emplois, ou la réalisation de l'ensemble de la R&D nécessaire pour développer ce modèle, devient rapidement extrêmement coûteux. | 3 recommandations :   * Les auteurs doivent indiquer et partager le temps de formation et sensibilité aux hyperparamètres :   Permettre de comparer directement différents modèles afin de réaliser une analyse coûts-avantages (précision).   * Les chercheurs universitaires ont besoin d'un accès équitable aux ressources informatiques : empêcher le monopole des grands groupes sur les algorithmes complexes pour permettre à plus de chercheurs d’étudier ces algorithmes et limiter la pollution de réentraîner des algorithmes simples ; * Les chercheurs doivent privilégier les hardwares et les algorithmes efficaces en termes de calcul :   Développer un algorithme capable de résoudre plusieurs problèmes, plutôt que d’implémenter un algorithme en réponse à chaque problème. |  |  |  |
| 8 | Comment quantifier les émissions de carbone de l'apprentissage automatique ?  Et développement d’un outil appelé le « Machine Learning Emissions Calculator » | * L’outil proposé par les auteurs prend en entrée plusieurs détails concernant l'entraînement du modèle de Machine Learning qui sont les suivants : le cloud utilisé, la zone géographique du serveur, le type de GPU utilisé ainsi que le temps d'entraînement de l'algorithmes. Après avoir analysé tous ces facteurs, l'outil donne en sortie la quantité approximative de CO2eq produite. Pour ces 4 variables, les chercheurs ont utilisé des données publiquement disponibles. * Afin de faire une estimation de la quantité de CO2eq émis par les serveurs cloud selon leur localisation. Les auteurs supposent que tous les serveurs sont connectés aux réseaux locaux à leur emplacement physique. Ainsi, des données concernant les émissions de CO2eq de différents emplacements de réseau ont été rassemblées et croisées avec les emplacements connus des serveurs GPU des trois principaux fournisseurs de cloud : Google Cloud Platform, Microsoft Azure et Amazon Web Services. |  | Les auteurs ont développé un outil capable de faire une approximation des émissions de CO2eq  4 recommandations :   * Choisir son cloud de façon intelligente * Choisir le lieu de son data center :   Bien que plusieurs fournisseurs de cloud soient neutres en carbone, certains de leurs data center peuvent tout de même émettre beaucoup de carbone dû au lieu où ils se trouvent. Il est donc très important de bien choisir le lieu du data center où l’algorithme s’entrainera.   * Limiter la perte de ressources :   Il est très important de faire les bons choix lorsque des tests et des expériences sont faites car ils peuvent être la raison d’énergies utilisées à perte.   * Bien choisir son Hardware :   Le choix du matériel peut aussi influer sur les émissions du Machine Learning. |  |  |  |
| 9 | Identifier les principaux facteurs qui contribuent à l'IA rouge et préconiser l'introduction d'une mesure d'efficacité simple et facile à calculer qui pourrait contribuer à rendre certaines recherches sur l'IA plus écologiques, plus inclusives et peut-être plus plausibles | * Montre que la plupart xdes recherches se sont basées sur la précision et non l’efficacité des programmes. Se focaliser sur l’amélioration de la précision est un facteur clé du red AI * EQUATION RED AI   Les auteurs ont identifié 3 pratiques qui contribuent au red AI   * Des modèles plus grands peuvent avoir des performances plus élevées, ce qui constitue une contribution scientifique précieuse. Toutefois, cela implique que le coût financier et environnemental de modèles d'IA de plus en plus grands ne diminuera pas de sitôt, car le rythme de croissance des modèles dépasse de loin l'augmentation de leurs performances qui en résulte. Par conséquent, de plus en plus de ressources seront nécessaires pour continuer à améliorer les modèles d'IA en les rendant simplement plus grands. * L'utilisation de données massives pour la formation crée des obstacles pour de nombreux chercheurs qui souhaitent reproduire les résultats de ces modèles ou former leurs propres modèles sur la même configuration. Enfin, comme dans le cas de la taille du modèle, s'appuyer sur davantage de données pour améliorer les performances est notoirement coûteux en raison du rendement décroissant de l'ajout de données supplémentaires. * Le thème du nombre massif d'expériences n'est pas aussi bien étudié que les deux premiers thèmes abordés ci-dessus. En fait, le nombre d'expériences réalisées pendant la construction du modèle est souvent sous-estimé. Néanmoins, il existe ici aussi des preuves d'une relation logarithmique entre le nombre d'expériences et les gains de performance.   D’après les auteurs pour mesurer l’efficacité il faut prendre en compte rendre disponible :   * + Les émissions carbones   + L’utilisation de l’électricité   + Le temps total pour générer une intelligence artificielle   + Le nombre de paramètres utilisés   + Le nombre de FPO * Pour démontrer l'importance de rendre compte de la quantité de travail, les auteurs présentent les coûts de la FPO pour plusieurs modèles existants et pour un seul modèle mais dont le nombre de couches varie | Les modèles deviennent plus coûteux avec le temps, mais l'augmentation du FPO ne conduit pas à des gains de performance similaires. Par exemple, une augmentation de près de 35 % du FPO entre ResNet et ResNext (deuxième et troisième point du graphique) a entraîné une amélioration de 0,5 % de la précision. Deuxièmement, le nombre de paramètres du modèle ne dit pas tout : AlexNet a en plus de paramètres que ResNet mais une précision beaucoup plus faible on peut voir cependant qu'il a aussi beaucoup moins de FPO. Dans la deuxième étude nous pouvons voir aussi qu’une forte augmentation des FPO ne se traduit pas par une aussi forte augmentation dans la précision |  | Le point sur les floating point n’est pas très clairs et les résultats ne concluent rien. On comprend qu’il faut prendre tout en compte mais la partie sur les floating point devrait être plus développée | Développer plus la partie des FOP et montrer l’impact des autres points évoqués par les auteurs | Faire le calcul des FPO avec notre algorithmes de classifications d’image en fonction des layers |

Note : toutes les études utilisent le CO2eq comme indicateur de l’impact environnemental et cela n’est pas suffisant car il ne prend pas tout en compte

Synthesis/Related work:

Prior academic research has emphasized the urgency of the situation and the gravity of the impacts of artificial intelligence, especially environmental. Indeed, numerous articles were written as call to action and realisation on that matter. One of the most cited literatures [Strubell et al. 2019] has shown the carbon emission issues and energy impacts of large NLP models by studying the power usage and carbon emissions for a set of case studies. The authors then made recommendations implying that “authors should report training time and sensitivity to hyperparameters”, “academic researchers need equitable access to computation resources”, and “researchers should prioritize computationally efficient hardware and algorithms”. These suggestions are shared and recognised by many. Then followed numerous recommendations such as the model of the 4 Ms [Patterson et al. 2022]. Here the authors argue that some of the best practices to reduce the energy consumption and the carbon emissions are “Model: selecting efficient ML model architectures”, “Machine: Using processors optimized for ML training”, “Mechanization: Computing in the cloud” and “Map: pick the location with the cleanest energy”. The quantification of the environmental impacts and finding a way to minimise them have become the subject and objective of several researchers who decided to broaden their view on the subject and to offer a holistic approach of the subject. [Acun et al. 2022] designed their “Carbon Explorer Framework” for carbon aware datacentres acknowledging the design space across renewable energy investments, energy storage, and computation shifting. [Wu et al. 2022] have also taken a holistic approach in exploring the environmental impacts of AI. The authors suggest that by optimizing different areas of LMs through judicious hardware-software such as “Platform-Level Caching”, “GPU acceleration” and “Algorithmic optimization” a 800× operational carbon footprint reduction for a Transformer-based universal language model.