INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN

Projet de Physique P6
STPI/P6/2024 – 02

IA Verte

Etudiants :

Sarah BENSLIMANE
Daniel VIARD
Kawtar EL GUEDDARI
Luiza MATTEDI
Audran LAUVERGEAT
Alexia DURAND

Enseignant-responsable du projet :

Abdelaziz BENSRHAIR

Date de remise du rapport : ***15/06/2024***

Référence du projet : ***STPI/P6/2024 – 02***

Intitulé du projet : ***IA Verte***

Type de projet : ***Bibliographie/ État de l'art***

Objectifs du projet :

Étude de l'intelligence artificielle verte à travers plusieurs axes :

- les IA classiques

- l'impact de l'intelligence artificielle sur l'environnement

- fondement de l'IA frugale

- solutions et application de l'IA Verte

Mots-clefs du projet : ***intelligence artificielle verte, environnement, impacts, solutions***

TABLE DES MATIERES

Index des figures

1. INTRODUCTION

L'intelligence artificielle (IA), stimulée par des avancées telles que les agents conversationnels comme ChatGPT, suscite un intérêt croissant à l'échelle mondiale. Ceci contribue de manière significative à l'augmentation de la consommation mondiale d'énergie dans le secteur numérique. Cette tendance préoccupante incite davantage de chercheurs à se concentrer sur la création d'une IA moins gourmande en énergie.

L'IA est le domaine de l'informatique qui se consacre au développement de systèmes capables d'effectuer des tâches qui requièrent généralement l'intelligence humaine. Ces systèmes sont capables de raisonner, d'apprendre, de percevoir et de planifier. L'IA a connu plusieurs avancées significatives depuis ses débuts dans les années 1950, avec des étapes clés telles que le développement des premiers réseaux de neurones dans les années 1980 et l'émergence de l'apprentissage profond dans les années 2010.

Dans ce contexte, l'objectif de ce rapport est d'étudier l'intelligence artificielle verte, nouveau domaine de l'intelligence artificielle consacré au développement et à la mise en œuvre de technologies respectueuses de l'environnement. Celui-ci traitera de l'IA verte en quatre grandes parties. Tout d'abord, nous présenterons les IA classiques et leurs domaines d'application, et discuterons de ses perspectives d'avenir et ses limites. Dans un second temps, nous nous pencherons sur les problèmes environnementaux que causent ces technologies. Nous aborderons, ensuite, l'IA frugale et son principe de base, discuterons de ses avantages et la comparerons aux IA classiques. Finalement, nous verrons les solutions qu'offrent l'IA frugale à l'aide d'une application concrète.

2. ORGANISATION DU TRAVAIL

Sarah Benslimane	***Kawtar El Gueddari***	***Alexia Durand***
IA classique/ rapport d'étonnement	*IA classique/ Introduction*	*Impact sur l'environnement*
Audran Lauvergeat	***Luiza Mattedi***	***Daniel Viard***
Fondement de l'IA frugale	*Fondement de l'IA frugale*	*Applications*

Répartition du travail par étudiant du groupe

Dans le cadre de ce projet, notre groupe a démontré ses capacités à travailler efficacement ensemble. Après avoir entamé nos recherches sur le sujet, établi un plan détaillé et des délais à respecter et désigné un chef de projet, nous avons pu rapidement nous répartir les tâches à accomplir. En fonction du travail demandé par chaque partie, une à deux personnes s'en sont chargées. Le reste du travail a été attribué sur la base du volontariat. Toutefois, nous avons décidé de réaliser la conclusion ensemble. Nous avons utilisé les séances hebdomadaires en classe avec monsieur Abdelaziz Bensrhair pour mettre en commun nos avancements et ajuster les délais fixés. Ces moments étaient essentiels pour le suivi du projet et pour s'assurer de la cohésion de nos parties. Nous avons veillé à ce que chaque partie du rapport s'enchaîne logiquement, formant ainsi une structure homogène et fluide.
Enfin, l'utilisation de document et de drive partagé ainsi que d'un groupe de discussion en ligne nous ont permis de rester en contact tout au long du semestre.
C'est ainsi que nous avons pu fournir un rapport complet et abouti pour le rendu final du projet.

3. TRAVAIL RÉALISÉ ET RÉSULTATS

3.1. Les IA classiques

3.1.1. Technologies existantes

L'intelligence artificielle englobe diverses méthodes et approches. Voici un aperçu des principales technologies actuelles :

Le Machine Learning: Connue aussi sous le nom d'apprentissage automatique, cette technique repose sur des algorithmes pour identifier des motifs récurrents dans les données. Ces données, qu'il s'agisse de chiffres, d'images ou de statistiques, sont stockées sous forme numérique. Les algorithmes de machine learning apprennent de manière autonome et améliorent leurs performances en découvrant ces motifs au fil du temps [1].

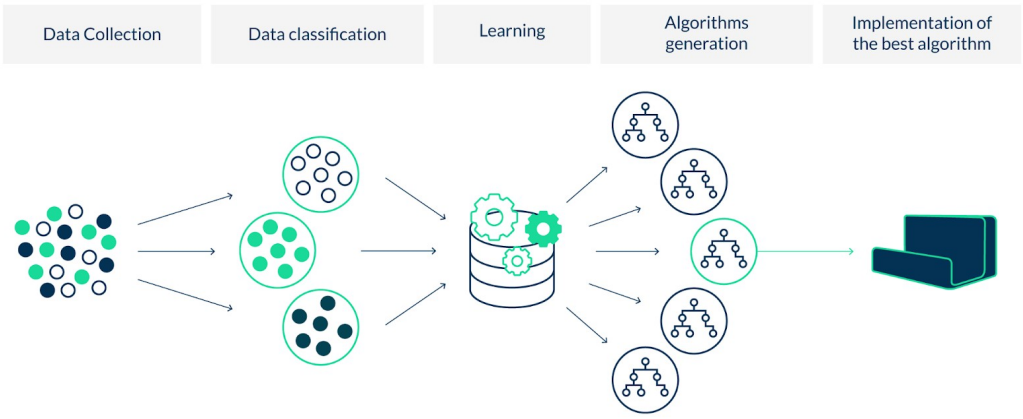

Figure 1: fonctionnement du machine learning

(Machine learning in IoT : did the dream come true? - PART II - Next4)

Le Deep Learning: Dérivé du machine learning, l'apprentissage profond permet à la machine d'apprendre par elle-même. Cette méthode repose sur un réseau de neurones artificiels qui imitent le cerveau humain. Chaque couche de neurones reçoit et interprète les informations de la couche précédente. Les réponses incorrectes sont éliminées et renvoyées en amont pour ajuster le modèle. Plus le système accumule d'expériences, plus il devient performant [2].

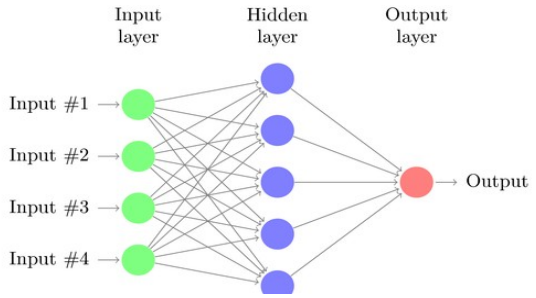

Figure 2: fonctionnement du deep learning

(Comprendre le deep learning – 2/3 : Fonctionnement (clevy.io))

Natural Language Processing: Le traitement automatique du langage naturel permet aux machines de comprendre, générer ou traduire le langage humain, écrit ou parlé. Un logiciel de NLP comporte trois composants : une interface visuelle, un moteur de traitement du langage naturel qui utilise un algorithme pour interpréter les requêtes des utilisateurs, et une console d'administration pour gérer les réponses à fournir [4].

Système expert: Cet outil peut répondre à des questions en se basant sur des faits et des règles établies. Un système expert est composé d'une base de faits, d'une base de règles et d'un moteur d'inférence capable de générer de nouveaux faits pour répondre aux questions posées [5].

Algorithmes de recherche heuristique: Ces algorithmes s'appuient sur des heuristiques, ou des règles pratiques, pour trouver une solution acceptable sans examiner exhaustivement toutes les options. Bien qu'ils soient rapides, ces algorithmes ne garantissent pas toujours la meilleure solution possible [6].

Systèmes de recommandation: Ce type de filtrage de l'information vise à proposer des éléments susceptibles d'intéresser l'utilisateur. En général, ces systèmes comparent le profil d'un client à des caractéristiques de référence et tentent de prédire les recommandations qu'un conseiller donnerait [7].

3.1.2. Les secteurs d'application de l'intelligence artificielle

Santé:
Dans le secteur médical, l'intelligence artificielle (IA) est cruciale pour les diagnostics assistés, les traitements personnalisés et la gestion des dossiers médicaux. Grâce à des algorithmes avancés, elle détecte des anomalies dans les images médicales, permet des interventions rapides et crée des plans de traitement adaptés au profil génétique des patients. Elle organise et sécurise également les informations médicales [8].

Finance:
L'IA révolutionne le secteur financier en détectant la fraude, en automatisant la gestion de portefeuille et en offrant des conseils d'investissement. Elle analyse les transactions pour identifier les activités suspectes et modernise les processus bancaires tout en facilitant les interactions avec les clients [9].

Education:
Dans l'éducation, l'IA propose des systèmes d'apprentissage personnalisés, l'évaluation automatique et des assistants virtuels. Elle adapte l'enseignement aux besoins des étudiants, libère les enseignants des tâches répétitives et fournit des informations sur les compétences des étudiants [10].

Industriel:
L'IA optimise la production industrielle, assure la maintenance prédictive et automatise le contrôle qualité. Elle anticipe les besoins en matériaux, prévient les pannes d'équipement et utilise la reconnaissance d'image pour détecter les défauts de fabrication [8].

Transport:
Dans les transports, l'IA permet le développement de véhicules autonomes, la gestion du trafic et l'optimisation logistique. Elle réduit les accidents, améliore l'efficacité des transports et diminue les embouteillages tout en optimisant la distribution des marchandises [8].

Ces applications de l'IA ne sont qu'un début, ouvrant continuellement de nouvelles opportunités et promettant un avenir où technologie et humanité collaborent pour résoudre des problèmes complexes.

3.1.3. Perspectives d'avenir

Le futur de l'intelligence artificielle s'annonce à la fois captivant et révolutionnaire. D'ici 2025, les entreprises prévoient d'accentuer leurs investissements dans la gouvernance des données et les plateformes d'IA pour passer à l'échelle de l'IA et du machine learning, conformément au rapport CIO Vision 2025 du MIT [11] . Celui-ci analyse les réponses de plus de 600 cadres de 18 pays différents en Amérique du Nord, Europe et Asie. 78% des sondés affirment que le développement de l'IA est leur principale priorité. L'impact prévu de l'IA d'ici 2025 s'étend à divers secteurs, notamment la santé, les transports et l'éducation.

La société américaine Gartner, spécialisée en conseil et recherche sur les technologies avancées, anticipe une adoption massive de l'intelligence artificielle dans le milieu professionnel, avec des outils d'IA soutenant les managers dans de nombreuses tâches [12]. Les entreprises mettent un accent majeur sur l'expansion de l'IA, en augmentant leurs investissements dans la gouvernance des données. En France, la stratégie nationale pour l'IA vise à encourager le développement de l'intelligence artificielle intégrée et à promouvoir son inclusion dans l'économie. La prolifération de l'IA aura un impact profond sur divers secteurs, notamment avec une grande partie du contenu en ligne générée par l'IA.

Les entreprises mettent une priorité cruciale sur l'intensification de l'IA, avec des investissements croissants dans la gouvernance des données. La stratégie nationale pour l'IA en France vise à promouvoir le développement de l'IA embarquée et favoriser son intégration dans l'économie. La diffusion massive de l'IA transformera radicalement divers secteurs, avec un contenu en ligne largement généré par l'IA.

Le futur de l'IA est susceptible d'être marqué par de nombreux progrès notamment le développement de supercalculateurs dédiés. Les supercalculateurs représentent des systèmes informatiques exceptionnellement puissants conçus pour exécuter des calculs complexes à des vitesses extrêmement élevées. Dans le contexte de l'IA, ces supercalculateurs spécialisés sont conçus pour accélérer les tâches d'apprentissage profond, de modélisation complexe et d'analyse de données massives. Grâce à leur capacité de traitement parallèle, ces machines peuvent manipuler simultanément d'énormes ensembles de données, accélérant ainsi les itérations d'entraînement des modèles d'IA.

L'un des supercalculateurs les plus récents et performants est le CS-3 de Cerebras Systems Inc. dont le processeur WSE-3 compte 4 trillions de transistors. Celui-ci a été conçu pour entraîner des modèles d'IA à plus de 24 millions de paramètres, jusqu'à 10 fois plus grand que Chat-GPT 4 [13].

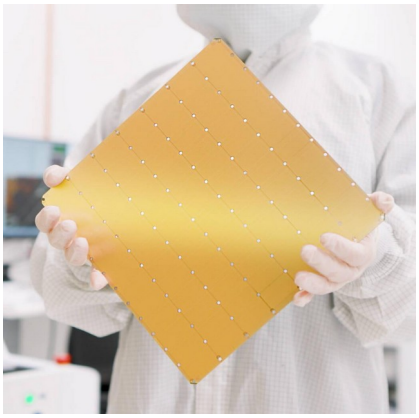

Figure 3: Wafer Scale Engine-3 © Cerebras

(https://www.futura-sciences.com/tech/actualites/technologie-plus-puissant-supercalculateur-taille-specialement-intelligence-artificielle-112211/)

On prévoit que les futurs supercalculateurs seront dotés d'architectures spécialisées, notamment des processeurs conçus spécifiquement pour des charges de travail liées à l'IA.

Ces avancées permettront d'atteindre des niveaux de performance inégalés, réduisant ainsi considérablement le temps nécessaire pour entraîner des modèles complexes et améliorant l'efficacité des applications d'IA à grande échelle. Ces dernières permettront des innovations dans des domaines tels que la recherche médicale, la découverte de médicaments, la simulation climatique, et bien d'autres.

Avec le développement rapide de solutions d'IA adaptées aux besoins des utilisateurs et l'émergence de plateformes simplifiées, de plus en plus de personnes ont la possibilité d'explorer et d'utiliser l'IA dans divers aspects de leur vie quotidienne. En augmentant l'accessibilité de l'IA au grand public, ceci ouvre ainsi la perspective d'une participation plus large et plus diversifiée à son développement.

3.1.4. Les limites de l'IA

Alors que les perspectives d'avenir de l'IA sont prometteuses, il est également crucial de reconnaître ses limites et ses défis inhérents. En effet, bien que les avancées technologiques puissent ouvrir de nouvelles possibilités, elles s'accompagnent souvent de questions éthiques, sociales et pratiques.

Tout d'abord, le développement de l'IA est limité par les ressources matérielles disponibles. De plus, les biais sont un problème majeur dans ces modèles. En effet, les bases de données utilisées peuvent être faussées ou pas assez représentatives. Dans le cas de la reconnaissance faciale, par exemple, les hommes blancs sont plus présents dans les banques d'images. L'algorithme s'entraîne d'une meilleure manière sur la population la plus représentée discriminant, ainsi, les autres [14].

Les modèles peuvent également manquer de robustesse face à des situations inattendues, faisant preuve d'une certaine fragilité. La compréhension limitée du contexte et l'incapacité à généraliser les connaissances à différentes situations représentent également des problèmes techniques majeurs. De plus, en raison de l'opacité des modèles d'IA, souvent due à leur complexité, il est difficile de comprendre et de tracer les décisions prises par ces systèmes.

Les questions éthiques liées à l'IA soulèvent des préoccupations fondamentales en matière de responsabilité, de respect de la vie privée et de justice sociale. L'utilisation de données personnelles dans le processus d'apprentissage des modèles soulève des questions éthiques liées à la protection de la vie privée et à la sécurité de l'information. Une fois des informations intégrées dans une IA, il est très difficile de supprimer cette information. Ceci pose problème dans le cas de données personnelles [15].

Ainsi, bien que les perspectives d'avenir de l'IA soient prometteuses, il est impératif de reconnaître et de surmonter les défis et les limites qui accompagnent son développement.

Après avoir examiné les différentes applications, avancées et limites des IA classiques, il est essentiel d'aborder un autre aspect de leur utilisation : leur impact sur l'environnement. En effet, alors que ces technologies continuent de se développer et de se déployer à grande échelle, il est de plus en plus nécessaire d'évaluer et de comprendre leur empreinte écologique.

3.2. L’impact de l’intelligence artificielle sur l’environnement

L'évolution rapide de l'intelligence artificielle (IA) a indéniablement transformé notre monde, apportant des avancées significatives dans divers domaines. Cependant, derrière les bénéfices évidents de l’IA se cachent des conséquences environnementales souvent négligées. Alors qu’elle continue de s'intégrer dans nos vies quotidiennes, son impact sur l'environnement devient une préoccupation de plus en plus importante et pourtant l’IA est une lueur d’espoir dans le rôle qu’elle peut avoir dans la préservation de l’environnement. Ainsi cette partie se penchera sur les problèmes causés par l’IA sur l’environnement, en examinant les différentes étapes du cycle de vie de l’IA, et ce qu’elle consomme concrètement.

Ainsi les intelligences artificielles ont un impact direct et un impact indirect sur l’environnement. L’impact direct est associée à l'intégralité du cycle de vie des dispositifs nécessaires à leur exécution, tels que les ordinateurs, serveurs, téléphones portables, cette catégorie englobe principalement la consommation énergétique de ces dispositifs lors de l'exécution des programmes, ainsi que la phase de fabrication et de fin de vie de ces équipements. L’impact indirect quant à lui découle de modifications dans d'autres secteurs ou de changements comportementaux, induits par l'utilisation de ces programmes [16].

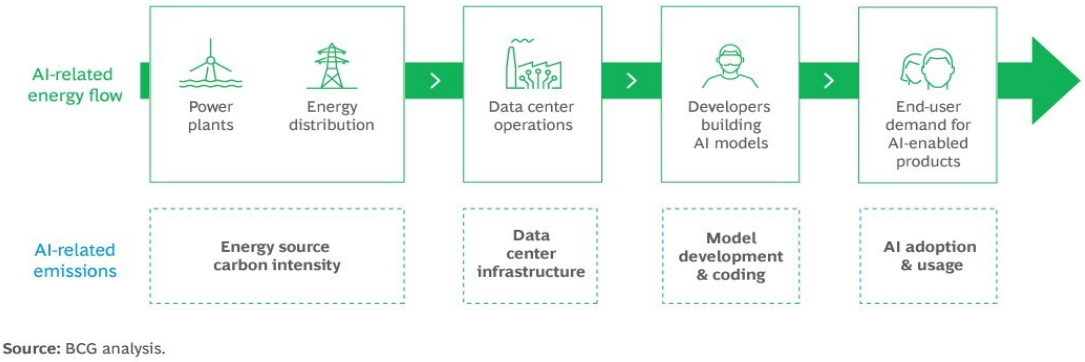

Figure 4 : cycle de vie de l’IA

(https://www.leptidigital.fr/intelligence-artificielle-ia/impact-ia-climat-51024/)

Le cycle de vie de l’IA se décompose en quatre parties :

- La phase de production intègre à la fois l'extraction physique des matières premières et la fabrication des composants essentiels pour construire le matériel et l'infrastructure de l'IA. L’empreinte carbone due à cette phase est infime.
- La seconde phase est le transport c’est-à-dire la distribution, le transport de marchandises, la manutention et le stockage du matériel informatique. L’empreinte carbone de cette seconde phase est également peu significative.
- La phase suivante concerne les opérations qui incluent la consommation en eau et en énergie. Cette phase est celle qui a le plus d’impact sur l’environnement. Bien que des améliorations d'efficacité aient été réalisées, l'entraînement de modèles d'IA

générative exige toujours une quantité significative d'énergie, avec des conséquences environnementales notables. De plus, la consommation d'eau pour la production d'électricité et le refroidissement des centres de données est également préoccupante.

- La dernière phase, la phase de fin de vie, inclut la collecte, l'envoie, le démantèlement, le recyclage, et l'élimination des déchets. L'impact environnemental majeur de la phase de fin de vie de l'IA réside dans sa contribution aux déchets électroniques. La désuétude des technologies de l'IA engendre des conséquences environnementales significatives en raison de la présence de matériaux nocifs, comme les métaux lourds et les produits chimiques toxiques, susceptibles de contaminer l'environnement [17].

Selon le rapport « The AI Disruption: Challenges and Guidance for Data Center Design » de Schneider Electric, la consommation électrique liée aux charges de travail d'IA devrait augmenter de manière significative d'ici 2028. Actuellement, elle se situe entre 4,3 GW, représentant environ 8% de la consommation totale des datacenters en 2023 (57 GW). En 2028, cette consommation devrait grimper à environ 14 à 19 GW, ce qui constituerait entre 15% et 20% de la consommation totale des datacenters, estimée à 90 GW à cette période [18]. L'analyse conclut que dans le domaine de l'IA, le processus d'apprentissage nécessite moins d'énergie que l'inférence. Cependant, il est important de noter que l'apprentissage lui-même demande une grande quantité de ressources, notamment pour le traitement de vastes ensembles de données[18].

En octobre 2023, une étude prépubliée réalisée par quatre chercheurs de Californie et du Texas prévoyait que d'ici 2027, les besoins en eau de toute l'industrie de l'IA pourraient être comparables à la moitié de la demande en eau du Royaume-Uni [19]. À titre d'exemple, une action en justice initiée par des habitants de Des Moines, la capitale de l'Iowa, a révélé que le centre de données d'OpenAI, où le dernier modèle GPT-4 était entraîné, avait utilisé 6 % de l'eau de l'ensemble du comté en juillet 2022 [20].

La gestion des déchets électroniques pose un défi majeur, surtout avec l'évolution rapide des équipements utilisés dans le domaine de l'IA. Cette évolution rend les appareils rapidement obsolètes et est souvent accélérée par les avancées technologiques constantes, ce qui contribue fortement à l'accumulation de déchets électroniques. Les composants spécifiques à l'IA, comme les GPU et les circuits intégrés dédiés, ajoutent une couche de complexité à cette problématique.

Quant à la surconsommation de ressources naturelles par l'IA, elle découle principalement de la demande croissante de matériaux nécessaires à la fabrication des équipements électroniques. Les métaux rares, par exemple, sont principalement utilisés dans la production de composants électroniques essentiels pour les technologies d'IA. Cette demande importante exerce une forte pression sur les ressources naturelles, entraînant des conséquences négatives telles que la déforestation, la pollution des sols et une diminution de la biodiversité.

3.3. Fondement de l'IA frugale

3.3.1. Principe de base et fonctionnement

L'IA frugale a un fonctionnement similaire à celui d'une IA classique, cependant il existe certaines différences pour la rendre plus écologique. Une IA frugale adopte plusieurs stratégies pour réduire les coûts, les besoins en ressources et la complexité tout en conservant des performances acceptables.

Une des premières méthodes est la sélection et/ou création de modèles légers : cela revient à utiliser des modèles plus simples et plus légers plutôt que d'utiliser des modèles d'IA complexes et gourmands en ressources. Cela peut impliquer l'utilisation de réseaux de neurones moins profonds, de réseaux de neurones à convolutions plus simples ou même de modèles linéaires.

Deuxièmement, on retrouve la compression de modèle qui consiste à réduire les temps de calculs et le stockage tout en économisant de l'énergie. La compression peut être réalisée de différentes manières. Tout d'abord par *quantification des poids: c*ela implique la réduction de la précision des valeurs, c'est-à-dire réduire le nombre de bits nécessaire lors d'une utilisation. Une autre technique serait le *pruning* qui consiste à réduire le nombre de connexions avec un faible poids. Le *pruning* permet donc de conserver uniquement les connexions nécessaires.

Après l'utilisation de cette technique, le modèle a besoin de moins de puissance de calcul et de stockage pour fonctionner. Après l'usage de ces deux techniques de compression, la complexité du modèle et les ressources nécessaires pour son fonctionnement diminuent.

Ensuite, le *transfert d'apprentissage (transfer learning),* une stratégie plutôt simple qui se traduit par l'usage de modèles pré-entraînés sur des ensembles de données massifs plutôt que par la formation d'un modèle à partir de zéro. Ces modèles peuvent être ensuite adaptés à des tâches spécifiques avec des ensembles de données plus petits et spécifiques. Cette stratégie permet principalement de gagner du temps, de l'énergie mais également d'améliorer les performances du modèle.

Dans le même principe que le *transfer learning*, l'optimisation du ré-entraînement des modèles consiste à créer une IA à partir d'un assemblage de plusieurs modèles plus petits pouvant apprendre indépendamment les uns des autres. Cette optimisation permet, tout comme le *transfer learning,* d'économiser de l'énergie et du temps.

L'IA frugale utilise également la méthode *on-device Inference.* En d'autres termes, plutôt que de faire tous les calculs sur des serveurs distants, les IA frugales effectuent autant de calculs que possible directement sur l'appareil ou le périphérique de l'utilisateur. Cela réduit la latence, économise de la bande passante et protège la confidentialité des données en ne transmettant que des résultats plutôt que des données brutes.

Enfin, la gestion des données: l'A frugale peut également être implémentée en utilisant des techniques de gestion des données, telles que l'échantillonnage intelligent des données pour réduire la taille des ensembles de données nécessaires à l'entraînement.

En résumé, L'IA frugale est une approche qui, à l'aide de diverses méthodes, permet , en sacrifiant une partie de sa complexité, d'obtenir des résultats satisfaisants [21].

3.3.2. Les avantages de l'IA frugale

L'IA frugale se démarque de l'IA traditionnelle par sa capacité à réaliser des tâches complexes avec des ressources limitées. Cette approche innovante offre des solutions efficaces tout en minimisant les coûts, ce qui en fait une alternative prometteuse dans le domaine de la technologie.

Sur le plan économique, l'IA verte possède de nombreux atouts, notamment en économisant du stockage. En utilisant moins de données, l'IA frugale permet de réduire significativement les coûts financiers en infrastructures et en énergie relatives à leur stockage. Ces économies ne sont pas négligeables, étant donné que le coût d'un datacenter, même de petite taille, s'élève en moyenne à 4,3 millions de dollars. Les datacenters peuvent être classés suivant 4 différents niveaux de 1 à 4. Les centres de niveau 1 sont moins coûteux et nécessitent moins d'énergie, mais sont également moins poussés avec un unique chemin de distribution de l'alimentation et du refroidissement sans composants redondants. Tandis que les centres de niveau 4 coûtant environ 15000€ le mètre carré, sont composés de plusieurs chemins de distribution de l'alimentation et de refroidissement actifs, et possèdent des composants redondants et sont tolérants aux pannes. Étant donné que l'IA verte utilise des méthodes permettant de réduire le nombre de données et la complexité du modèle, l'usage d'un centre de données plus petit pour une IA frugale est plus adapté et bien moins coûteux.

De même pour réduire les coûts, L'IA verte s'équipe de composants plus économes sur le plan financier mais également écologique. En effet, depuis l'augmentation des prix des composants due à la pandémie COVID-19, le fait que certains processeurs graphiques (GPU) atteignant jusqu'à 2 500 dollars pièce, et la volatilité élevée des coûts de l'énergie, les entreprises sont confrontées à des défis financiers considérables. C'est dans ce contexte que les composants conçus pour offrir à la fois frugalité et efficacité deviennent un atout stratégique majeur. Parmi ces composants, les puces neuromorphiques se distinguent, s'inspirant du fonctionnement des neurones et des synapses biologiques. Leur attrait réside dans leur capacité à consommer jusqu'à mille fois moins d'énergie qu'un processeur conventionnel de même taille, en réduisant notamment les transferts de données.

Finalement, en minimisant les coûts d'annotation, différentes méthodes peuvent être utilisées pour les réduire, tel que le *few shot learning* dont le principe est la réduction du jeu de données ou encore l'automatisation de l'annotation pouvant permettre d'éviter l'annotation de gros jeux de données, une étape coûteuse en ressources financières et humaines.

C'est sur le plan écologique que l'IA frugale prend tout son sens, tout d'abord grâce à l'usage de différentes méthodes telles que le *transfer learning*, le *on device inference* qui permettent pour l'une d'éviter l'étape d'entraînement de l'IA, une étape extrêmement énergivore, et pour l'autre de réaliser un maximum de calculs sur l'appareil de l'utilisateur plutôt que de les

effectuer sur le cloud qui consomme 5,15 Mwh d'électricité par m² /an pour un centre de données situé en France. De même, en optimisant le ré-entraînement des modèles, l'IA frugale est conçue avec une architecture qui minimise la phase énergivore de l'apprentissage.

Comme mentionné dans les avantages économiques, une IA verte peut se contenter d'un data center de petite taille, le bénéfice n'est pas uniquement économique mais également écologique. En effet, en 2022 , les 250 data centers situés en France consomment en moyenne 5,15 MWh m²/an. D'après l'étude menée en 2022 par l'ADEME et l'ARCEP, les centres de données représentent 2,7% de l'empreinte carbone française. Selon une étude de l'ADEME réalisée en 2022, les datas centers sont à l'origine de 3 à 4% des émissions de gaz à effets de serre dans le monde. Ainsi, plus un data center sera petit, moins il consommera et par conséquent diminuera nettement les émissions dues à l'usage de l'électricité et les émissions de gaz à effets de serre indirect. En dépit d'une consommation déjà élevée d'électricité, les data centers consomment aussi un quantité d'eau exorbitante de 600 000 mètre cubes d'eau par an afin de refroidir les serveurs stockant les données dans les data centers. Une fois de plus, la conclusion est la même que pour la consommation d'électricité, la possibilité de se contenter d'un data center de petite taille permet de réduire de manière conséquente la consommation d'eau par rapport à une IA classique.

Finalement, L'IA frugale s'avère être précieuse pour limiter notre empreinte environnementale en réduisant sa consommation d'énergie, cela grâce aux différentes méthodes utilisées pour réduire son stockage et sa complexité tout en garantissant une vitesse de calcul et des résultats satisfaisants.

3.3.3. Comparaison avec les IA classiques

Les IA vertes (éco-responsables) et les IA classiques présentent des différences en termes de leurs caractéristiques et de leurs impacts environnementaux.

Pour les intelligences artificielles classiques, la consommation énergétique est très élevée. Cela vient du fait qu'elles sont souvent basées sur des réseaux neuronaux profonds et des modèles complexes, qui nécessitent d'importantes ressources de calcul. De plus, ces IA sont souvent exécutées sur des serveurs à grande échelle, qui consomment beaucoup d'électricité pour fonctionner et pour refroidir les systèmes.

Selon le rapport "The AI Disruption: Challenges and Guidance for Data Center Design"[22] de Schneider Electric, les IAs ont non seulement consommé 4,5 GW en 2023, mais ont aussi une prédiction de consommer environ 16 GW en 2028.

Ensuite, les modèles actuels traitent énormément de données, car les IA classiques ont tendance à nécessiter de grandes quantités de données pour l'apprentissage et l'entraînement de modèles. Cela a pour conséquences des transferts de données massifs à travers les réseaux, ce qui peut augmenter la consommation d'énergie.

Enfin, l'impact environnemental est fort, puisque ces modèles peuvent avoir des émissions de carbone élevées, contribuant ainsi aux changements climatiques. Ces émissions sont dues en partie à la réfrigération des data centers, qui eux contribuent à 2% des émissions de gaz à effet de serre (GES) mondiales [23].

En revanche, les intelligences artificielles vertes proposent une optimisation de la consommation énergétique. Elles visent à minimiser la consommation d'énergie en ne traitant que les données pertinentes lorsqu'elles sont nécessaires.

Ces IA se concentrent sur l'utilisation efficace de l'énergie en adoptant des approches telles que le calcul événementiel. Au lieu de traiter en continu de vastes quantités de données, l'IA verte se concentre sur les changements significatifs, ce qui réduit la nécessité de calculs constants. Cela permet une réduction significative de la consommation d'énergie lors des périodes d'inactivité ou de faible activité.

En outre, elles ont besoin de moins de données, car les IA vertes cherchent à réduire la dépendance aux grands ensembles de données pour l'apprentissage en se concentrant sur l'apprentissage à partir de données en temps réel ou locales.

Finalement, les impacts environnementaux seront réduits, du fait qu’en réduisant la consommation d'énergie et en minimisant le besoin de grands serveurs de calcul, les IA vertes ont un impact moindre sur l'environnement.

En comparaison, les IA classiques ont tendance à être plus gourmandes en énergie, à nécessiter plus de ressources de calcul et à avoir des impacts environnementaux plus importants.

Plus récemment, en mars 2024, l’entreprise Nvidia, leader dans le calcul IA, a façonné une nouvelle super puce pour IA qui pourrait aider à diminuer la consommation d’énergie de ces technologies. Cette puce se nomme Blackwell B200, un nouveau GPU moins énergivore. Cet avancement permettra aux entreprises d’utiliser des puissances IAs avec moins de consommation d’énergie.

3.4. Solutions de l'IA Verte

Depuis 2022 et l'essor des IA génératives textuelles tel que ChatGPT, de nombreuses multinationales spécialisées dans la Tech investissent dans ce domaine et créent leurs propres variantes. L'entreprise Apple, leader sur le marché du smartphone, ne déroge pas en y investissant près d'un milliard de dollars américains par an [24]. Plus de 24 entreprises spécialisées ont également été reprises par la pomme [25]. Pourtant en suivant les annonces de la firme (à l'image de l'Apple Worldwide Developers Conference 2023), le terme IA n'y est jamais évoqué. La raison ? Apple ne s'oriente pas en priorité vers les IA génératives mais vers les IA intuitives, rendant l'expérience plus fluide et donnant à l'utilisateur des outils pratiques et non inclusifs [26]. En faisant ce choix, Apple cherche à réduire la puissance de calcul demandée, et s'oriente ainsi vers un modèle d'IA verte. L'IA générative en est alors qu'en développement, car encore non implantable exclusivement on-device, c'est-à-dire en se passant des interactions avec des data centers. Ainsi, analysons précisément la situation de cette entreprise, en comprenant tout d'abord les notions d'IA génératives et intuitives et comment ces dernières s'immiscent dans la vision du futur du PDG Tim Cook. Nous pourrons expliciter la situation actuelle du marché du smartphone pour en comprendre les enjeux dans le domaine de l'intelligence artificielle et finir par les méthodes d'implantation des IA et IA frugales utilisées par Apple pour répondre à ces problématiques si importantes.

3.4.1. La place de l'IA intuitive dans la vision d'Apple

Commençons par un court point sur les IA génératives et intuitives [27]. Ces deux types se basent sur le deep learning pour répondre de manière adaptée à une situation, reproduisant des schémas de pensée humains. La différence réside dans la spécification du problème et le but recherché. Par exemple, aux échecs, une IA intuitive s'entraîne et analyse des parties pour déterminer le meilleur coup. En revanche, une IA générative, avec plus de données, peut créer de nouvelles stratégies nécessitant plus de calculs. L'IA intuitive résout des problèmes précis et prend des décisions, tandis que l'IA générative crée du contenu Human-Like.

L'IA générative, en s'inspirant de millions de situations, excelle dans la création pure, comme les images de MidJourney, et répond rapidement à de nombreuses questions. Son adaptabilité est une force, mais aussi une faiblesse, car elle manque de recul sur les informations incorrectes, nécessitant une vérification. La précision est donc limitée.

L'IA intuitive apprend de données centrées sur une démarche précise, réduisant les erreurs et optimisant la tâche. Elle est utile pour l'automatisation de tâches et la prédiction, comme en analyse de données business, étant plus précise que les analyses humaines. Ce type d'IA est spécialisé et moins adaptable.

La vision d'Apple combine le monde réel et virtuel, à l'image du Vision Pro [28], un casque de réalité virtuelle permettant d'effectuer toutes les tâches d'un ordinateur sans périphériques, grâce à la détection des mouvements des yeux et des mains. L'IA intuitive est utilisée pour la détection, les opérations sur l'image et l'environnement, et la traduction en temps réel. Le défi est de transmettre cela sur les smartphones avec une puissance de

calcul limitée, tout en traitant des millions de données personnelles. La précision limitée des IA génératives, nécessitant de grandes banques de données et une puissance de calcul importante, pousse Apple vers l'IA intuitive, en accord avec sa réputation de fiabilité et de qualité. L'IA générative pourrait arriver dans le futur, mais les défis sont nombreux.

3.4.2. Les défis de l'IA sur le marché du Smartphone

Aujourd'hui, l'IA est un élément clé du marketing de chaque entreprise, y compris pour les smartphones, souvent appelés AI phones [29], mais les limites, selon Apple, semblent encore trop importantes. Quelles sont ces limites et sont-elles réellement insurmontables ? Et pourquoi d'autres entreprises, comme Samsung, prétendent les avoir franchies [30] ?

La démocratisation du online permettrait d'utiliser la puissance des data centers, mais cela soulève des problèmes de sécurité et d'impact environnemental. La sécurité est un frein majeur, comme l'a montré une récente faille de sécurité chez France Travail [31]. Réduire les flux de données sensibles semble donc crucial. Il est par ailleurs amusant de noter que les IA on-device, comme la reconnaissance faciale, sont utilisées depuis de nombreuses années pour renforcer la sécurité des smartphones.

L'entraînement de ChatGPT-3 [32] nécessite énormément de ressources, avec une consommation d'eau et d'énergie considérable. Les applications mobiles online contribuent également à l'empreinte carbone, émettant jusqu'à 10g de CO_2 par jour et par personne, selon certaines estimations [33]. Les IA génératives, comme celle de Snapchat, augmentent également cette empreinte. En tête des applications les plus polluantes [34], TikTok, Reddit, Youtube et Instagram ont un impact considérable avec près de 5 % des émissions mondiales de gaz à effet de serre qui sont attribuées à ces applications, sans compter le rechargement des mobiles.

3.4.3. L'IA frugale au sein des Smartphones

Actuellement, environ 35% des smartphones intègrent des fonctionnalités basées sur l'IA [35], telles que les retouches photos, les assistants personnels, la sécurisation de l'appareil et les recherches sur internet. Pour répondre à ces demandes croissantes, les AI-phones utilisent des NPUs (Neural Processing Units), des microprocesseurs dédiés à l'IA, séparés des CPU et GPU traditionnels [36].

Apple et Huawei sont en tête dans l'utilisation de ces NPUs, permettant une exécution efficace des tâches basées sur l'IA on-device. Par exemple, l'iPhone X de 2017 disposait d'un ANE (Apple Neural Processor) avec un débit maximal de 0,6 téraflops, tandis que l'iPhone 13 atteint jusqu'à 15,8 téraflops. Cette évolution des NPUs individuels est encourageante pour l'avenir de l'IA on-device, d'autant plus que les performances des nouvelles puces NVIDIA sont impressionnantes [37].

La rationalisation de l'IA générative, comme le fait Apple avec Siri, est également une approche pour réduire les échanges online. Il est nécessaire de limiter l'utilisation de l'IA générative par les applications tierces pour des raisons de sécurité et d'impact environnemental, ce qu'on peut attendre d'Apple via son Apple Store déjà restrictif.

Apple se distingue enfin par le contrôle par l'IA de la batterie pour optimiser sa durée de vie et ses rechargements. Des ajustements automatiques des paramètres du smartphone peuvent même être suggérés pour économiser de l'énergie [38].

À l'avenir, on peut imaginer des puces et réglages spécifiques développés pour répondre aux besoins variés des utilisateurs, comme ceux axés sur le gaming, la bureautique ou la photographie, tandis que les produits plus utilitaires pourraient limiter l'usage de l'IA.

Pour conclure, le marché du smartphone est hautement concurrentiel, avec un fort accent sur le développement de l'IA pour des raisons marketing. Les fonctionnalités basées sur l'IA intuitive sont déjà répandues, tandis que l'IA générative émerge, posant des défis environnementaux et de sécurité liés à son entraînement et à son fonctionnement online. Bien que le développement des processeurs dédiés à l'IA soit prometteur, il ne résout pas complètement ces problèmes. Une rationalisation des usages, un contrôle des applications et une sensibilisation sont nécessaires.

Apple se démarque en n'intégrant pas d'IA générative tant que la technologie ne le permet pas, en contrôlant les performances de leurs appareils et en étant à la pointe des avancées dans ce domaine. Cependant, il est crucial de reconnaître que l'impact global du marché du smartphone, y compris celui d'Apple, est désastreux sur le plan environnemental et social.

4. CONCLUSIONS ET PERSPECTIVES

Ce projet met en lumière l'importance du développement d'une IA verte. Dans un premier temps, nous avons pu mieux nous familiariser avec les modèles d'IA classiques en étudiant leurs secteurs d'applications, leurs perspectives d'avenir et leurs limites. C'est ainsi que nous avons introduit les problèmes environnementaux que posent ces modèles tels que la surconsommation de ressources naturelles et énergétiques ou la gestion des déchets électroniques. Nous avons ensuite abordé le principe et les avantages de l'IA frugale qui en plus d'être plus écologique que les IA classiques, permet notamment de minimiser le stockage, les coûts financiers et humains. Enfin, nous avons illustré le projet par la stratégie d'Apple en matière d'IA, soulignant ses efforts pour optimiser l'utilisation de l'IA tout en minimisant ses impacts négatifs.

Ce projet nous a apporté d'un point de vue personnel de nouvelles connaissances mais également une approche de travail différente. Tout d'abord, pour réaliser ce projet nous avons dû travailler en groupe de 6. Il a donc fallu que l'on s'organise et se répartisse le travail afin de pouvoir avancer efficacement entre chaque séance. Ainsi ce projet nous a apporté une expérience supplémentaire et qui sera bénéfique pour nos futures expériences professionnelles. D'autre part, ce projet nous a enrichi intellectuellement, puisque nous nous sommes intéressés de près à l'intelligence artificielle qui a une place de plus en plus significative dans nos vies. Nous avons également pu nous rendre compte de l'impact environnemental de l'IA et avons pris connaissance de l'IA frugale. Ainsi nous ressortons de ce projet enrichi sur plusieurs aspects.

Pour approfondir ce sujet large et de fait complexe, nous pourrions nous concentrer sur l'avancement de l'implémentation des principes d'IA frugale au sein des différents domaines, et de la prise en compte de ces derniers par les entreprises dites "tech" grand public. De même, il serait intéressant d'étudier les démarches réelles réalisées dans le but de faire prendre conscience de la situation et des implications écologiques de l'utilisation de l'IA. Enfin, les limites de l'IA frugale devraient être analysées précisément afin de définir les domaines pouvant et ne pouvant pas s'y conformer.

5. REMERCIEMENTS

Nous tenons à exprimer notre gratitude à toutes les personnes et institutions qui ont contribué à la réalisation de ce projet scientifique au sein de l'INSA.

Un grand merci à notre encadrant, Abdelaziz Bensrhair , pour son soutien et ses conseils avisés. Nous remercions également le corps professoral de l'INSA et les membres du laboratoire LITIS pour leur aide technique et leurs ressources.

6. RAPPORT D'ÉTONNEMENT

L'intelligence artificielle a parcouru un chemin impressionnant ces dernières années, dépassant les attentes et repoussant les limites de ce que nous pensions possible. Au milieu de cet élan d'innovation, émerge un concept intrigant et novateur : l'IA frugale. Au cours de notre projet, nous avons étudié en détail ce concept. Dans ce qui suit, nous rappellerons brièvement les aspects écologiques en relation avec l'IA. Néanmoins, nous aborderons également les aspects sociaux, économiques et éthiques de l'IA.

L'impact environnemental de l'IA est souvent négligé, mais il est important. Les modèles d'IA traditionnels sont souvent volumineux et complexes, et nécessitent de grandes quantités de ressources énergétiques pour être entraînés et exécutés. Cela augmente considérablement les émissions de dioxyde de carbone, contribuant ainsi au changement climatique et à la dégradation de l'environnement. La réponse à ce problème est l'IA verte, également connue sous le nom d'IA frugale.

Sur le plan social, l'IA verte présente plusieurs avantages. En rendant l'IA plus respectueuse de l'environnement, elle contribue à protéger les ressources naturelles et à lutter contre le changement climatique, au bénéfice de la société dans son ensemble. De plus, en réduisant les coûts associés à l'utilisation de l'IA, elle peut favoriser une adoption plus large et démocratiser l'utilisation de la technologie, créant ainsi des opportunités pour les populations marginalisées.

D'un point de vue économique, l'IA écologique offre des perspectives intéressantes. En réduisant la consommation d'énergie et les coûts d'exploitation, les entreprises et les organisations peuvent augmenter leur rentabilité tout en réduisant leur empreinte environnementale. De plus, en encourageant l'innovation dans le secteur des technologies propres, elle peut stimuler la croissance économique tout en contribuant à la transition vers une économie plus durable.

Enfin, d'un point de vue éthique, l'IA verte soulève d'importantes questions sur la responsabilité sociale et environnementale des entreprises. Il est essentiel de garantir que les solutions d'IA verte sont développées et déployées de manière éthique, en tenant compte des impacts sociaux et environnementaux à long terme. Cela implique de veiller à ce que les avantages de la technologie soient équitablement répartis et que les risques environnementaux soient minimisés.

En résumé, l'IA verte représente une évolution importante dans le domaine de l'intelligence artificielle, offrant des solutions plus respectueuses de l'environnement, plus rentables et socialement responsables. Afin de maximiser ses bénéfices et de réduire ses risques, les dimensions écologiques, sociales, économiques et éthiques de cette approche doivent être prises en compte. En investissant dans la recherche et le développement en IA verte de manière responsable et inclusive, nous pouvons réaliser tout son potentiel et construire un avenir où technologie et durabilité vont de pair.

7. BIBLIOGRAPHIE

[1] : Machine Learning : Définition, fonctionnement, utilisations (datascientest.com)consulté le 16/03/24
[2] : Définition | Deep Learning - Apprentissage profond (futura-sciences.com) consulté le 16/03/24
[3] : Le traitement du langage naturel (NLP), comment ça marche ? | dydu consulté le 10/03/24
[4] : Natural language processing (NLP) : définition et techniques (journaldunet.fr) consulté le 18/03/24
[5] : Système expert — Wikipédia (wikipedia.org) consulté le 18/03/24
[6] : Qu'est-ce qu'un algorithme de recherche heuristique ? ▷➡ (tecnobits.com) consulté le 18/03/24
[7] : Système de recommandation – Wikiwand consulté le 18/03/24
[8] : L'intelligence artificielle et ses domaines d'application (litiliste.com) consulté le 6/04/24
[9] : Qu'est-ce que l'intelligence artificielle dans la finance ? | IBM consulté le 6/04/24
[10] :www.researchgate.net/publication/346952655_L'intelligence_artificielle_en_education_un_apercu_des_possibilites_et_des_enjeux consulté le 6/04/24
[11] : www.databricks.com/wp-content/uploads/2022/09/cio-vision-2025-final.pdf consulté le 10/03/24
[12] :www.lemondeinformatique.fr/actualites/lire-l-ia-au-travail-une-diffusion-massive-en-2025-76446.html consulté le 12/03/24
[13] :www.futura-sciences.com/tech/actualites/technologie-plus-puissant-supercalculateur-taille-specialement-intelligence-artificielle-112211/ consulté le 8/04/24
[14] :www.ellisphere.com/les-biais-caches-de-lia/ consulté le 20/04/24
[15] :secureprivacy.ai/fr/blog/ia-protection-des-donnees-personnelles-conformite-au-gdpr-et-au-ccpa consulté le 20/04/24
[16] :ecoinfo.cnrs.fr/2019/10/01/impact-environnemental-de-lia/ consulté le 20/03/24
[17] :www.schroders.com/fr-fr/fr/particuliers/paroles-d-experts/revolution-de-l-ia-quel-impactenvironnemental/ consulté le 20/03/24
[18] :dcmag.fr/la-consommation-energetique-de-lia/ consulté le 20/03/24
[19] :arxiv.org/abs/2304.03271 consulté le 20/03/24
[20] :www.la-liberte.ca/2024/03/05/lempreinte-environnementale-sous-estimee-de-lia/ consulté le 20/03/24
[21] :france-science.com/adopter-lia-frugale-concepts-leviers-et-initiatives/ consulté le 20/03/24
[22] : "The AI Disruption: Challenges and Guidance for Data Center Design", de Schneider Electric.
[23] :www.actu-juridique.fr/administratif/seine-saint-denis-et-data-centers-la-grande-invasion/ consulté le 25/03/24
[24] : CNBC - apple spend 1 billion a year in ai / Bloomberg - how apple plan to bring ai in his devices consulté le 25/03/24
[25] : MacGeneration - apple rachète des stratups dans l'IA consulté le 25/03/24
[26] : Wired - intuitive ai for Iphone 15 / Wired - Apple ghost generative ai consulté le 26/03/24
[27] : mynoec - intuitive ai vs generative ai / eweek - generative ai vs ai consulté le 10/04/24
[28] : apple : vision pro consulté le 10/04/24
[29] : CNBC : AI phones consulté le 10/04/24
[30] : CNBC : new samsung S24 ai features consulté le 10/04/24
[31] : OuestFrance - attaque sur pôle emploi. consulté le 10/04/24
[32] : lebigdata - le vrai cout de chat gpt / alliancy - emprreinte carbone de chat gpt consulté le 13/04/24
[33] : leparisien - applications et pollution passive. consulté le 21/04/24
[34] : greenspector - empreinte environnementale des réseaux sociaux consulté le 21/04/24
[35] : greenbook - how is ai transforming mobile tech consulté le 21/04/24
[36] : machinelearning.apple - Apple Neural Engine consulté le 21/04/24
[37] : nvidia - DGX B200 consulté le 21/04/24
[38] : apple - Optimized Battery Charging / linkedin - hidden ai features on Iphone consulté le 21/04/24