PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
PRÓ-REITORIA DE GRADUAÇÃO
ESCOLA POLITÉCNICA E DE ARTES
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

DESAFIOS ENERGÉTICOS EM TREINAMENTO DE MODELOS DE INTELIGÊNCIA ARTIFICIAL

VITOR FRANÇA

GOIÂNIA-GO
2024

VITOR FRANÇA

DESAFIOS ENERGÉTICOS EM TREINAMENTO DE MODELOS DE IA

Trabalho de Conclusão de Curso apresentado na escola Politécnica e de artes da Pontifícia Universidade Católica de Goiás como requisito básico para a conclusão do curso de Ciência da Computação.
Prof. Orientador: Prof. Me. Gustavo Siqueira Vinhal.

GOIÂNIA/GO

2024

VITOR FRANÇA

DESAFIOS ENERGÉTICOS EM TREINAMENTO DE MODELOS DE IA

BANCA EXAMINADORA

_____

Orientador: Prof. Me. Gustavo Siqueira Vinhal

_____

Examinador Convidado : Prof. Me. Carlos Alexandre Ferreira de Lima

_____

Examinador Convidado: Prof. Me. Rafael Leal Martins

SUMÁRIO

RESUMO

Este trabalho aborda os desafios energéticos gerados pelo avanço da tecnologia de Inteligência Artificial (IA). Com o aumento exponencial no uso de IA e a crescente complexidade dos modelos, o consumo energético tornou-se uma preocupação significativa. Destaca-se a importância de explorar técnicas e práticas para otimizar o consumo de energia, incluindo o uso de algoritmos mais eficientes e a implementação de *data centers* sustentáveis. Além disso, o trabalho aborda a possibilidade de adotar novas tecnologias com o potencial promisso, como o disco óptico 3D em escala nanométrica com capacidade de *petabit*, que podem reduzir o impacto ambiental do treinamento de IA, sem comprometer a eficácia dos modelos. Diante disso, o presente trabalho tem por objetivo refletir acerca do impacto ambiental da IA no mundo e fornecer *insights* valiosos para a comunidade acadêmica e profissionais da área, incentivando a adoção de práticas sustentáveis no desenvolvimento de tecnologias de IA buscando soluções que permitam a utilização dessa tecnologia de forma consciente e sustentável.

Palavras-chave: Inteligência Artificial; consumo energético; impacto ambiental; sustentabilidade; otimização energética; disco óptico 3D;

ABSTRACT

This paper addresses the energy challenges posed by the advancement of Artificial Intelligence (AI) technology. With the exponential increase in AI usage and the growing complexity of models, energy consumption has become a significant concern. This study highlights the importance of exploring techniques and practices to optimize energy consumption, including the use of more efficient algorithms and the implementation of sustainable data centers. Additionally, it discusses the potential of new technologies, such as nanometric-scale 3D optical disks with *petabit* capacity, to reduce the environmental impact of AI training without compromising model efficiency. The primary objective of this work is to reflect on the environmental impact of AI globally and provide valuable insights to the academic community and professionals in the field, encouraging the adoption of sustainable practices in the development of AI technologies. The specific objectives include: identifying the main energy challenges in AI model training, evaluating current techniques and practices for energy optimization, highlighting promising new research and methods for energy saving, and proposing appropriate policies and regulatory practices to ensure a balance between performance and sustainability.

Keywords: Artificial Intelligence; energy consumption; environmental impact; sustainability; energy optimization; 3D optical disc;

Lista de Siglas

- IA - Inteligência Artificial
- ML - *Machine Learning* (Aprendizado de Máquina)
- LLM - *Large Language Models* (Grandes Modelos de Linguagem)
- GEE - Gases de Efeito Estufa
- CO_2 - Dióxido de Carbono
- TPUs - Unidades de Processamento Tensorial
- GPUs - Unidades de Processamento Gráfico
- CNNs - Redes Neurais Convolucionais
- RNNs - Redes Neurais Recorrentes
- GDPR - General Data Protection Regulation (Regulamento Geral sobre a Proteção de Dados)
- LGPD - Lei Geral de Proteção de Dados
- HPS - Hexafenilsilol
- AIE - Emissão Induzida por Agregação
- ITX - Isopropil tioxantona (fotoiniciador)
- DTPA - Ácido Dietilenotriaminopentacético
- *AIEgens* - Luminógenos com característica de emissão induzida por agregação

INTRODUÇÃO

O desenvolvimento da inteligência artificial (IA) é uma das transformações tecnológicas mais impactantes do século XXI, trazendo consigo uma série de desafios e oportunidades. O treinamento de modelos de IA em nuvem, em particular, destaca-se pela sua capacidade de processar enormes quantidades de dados e aprender com eles, o que é essencial para a criação de modelos robustos e eficientes. Esta tecnologia tem aplicações abrangentes em setores como saúde, finanças, transporte e educação, proporcionando avanços significativos, tais como diagnósticos médicos mais precisos, sistemas de transporte mais seguros e processos financeiros mais eficientes. No entanto, a crescente demanda por processamento em nuvem implica em um consumo energético considerável, o que levanta preocupações ambientais e de sustentabilidade.

A influência positiva da IA na sociedade é indiscutível. Ao automatizar tarefas comuns e repetitivas, a IA aumenta a eficiência e a produtividade em diversos setores. No entanto, a infraestrutura necessária para suportar o treinamento e desenvolvimento de modelos de IA consome uma grande quantidade de energia, não apenas para processamento, mas também para refrigeração e manutenção. Nesse contexto, a implementação de *data centers* projetados para eficiência energética é crucial. O uso de fontes de energia renovável e sistemas de armazenamento inovadores pode contribuir significativamente para a redução do consumo energético, beneficiando a todos.

O objetivo geral deste trabalho é examinar os desafios energéticos associados ao treinamento de modelos de IA, visando identificar os principais obstáculos e propor soluções para otimizar o uso de energia, equilibrando desempenho e sustentabilidade. Este estudo investiga as práticas atuais de consumo energético, tecnologias emergentes voltadas para a otimização energética e políticas regulatórias que possam impactar positivamente a eficiência energética. Assim, busca-se contribuir para a redução do impacto ambiental e promover práticas sustentáveis no desenvolvimento e uso de modelos de IA.

Para atingir esse objetivo, o trabalho se concentra nos seguintes aspectos específicos: (1) identificar os principais impactos ambientais do treinamento de IA, além dos desafio e consumo energéticos envolvidos no treinamento de modelos de IA; (2) avaliar as estratégias e práticas atuais para otimização do consumo

energético, incluindo o uso de algoritmos mais eficientes e a implementação de *hardware* específico e implementação de computação em nuvem; (3) explorar novas linhas de pesquisa e tecnologias com potencial para economia de energia no treinamento de IA; (4) analisar políticas e práticas regulatórias que assegurem um equilíbrio entre desempenho e sustentabilidade, promovendo a eficiência energética e minimizando o impacto ambiental associado ao treinamento de modelos de IA.

A metodologia deste trabalho é baseada em uma análise crítica das práticas atuais e emergentes de otimização energética no contexto do treinamento de IA, juntamente com uma avaliação de estudos que destacam as novas práticas e tecnologias promissoras para a economia de energia.

Diante desse cenário, com esse estudo, espera-se fornecer uma análise abrangente dos desafios energéticos no treinamento de modelos de IA em nuvem, oferecendo recomendações práticas e uma visão de possíveis soluções para promover a sustentabilidade. Entender esses aspectos é essencial para orientar políticas públicas, regulamentos e práticas éticas no desenvolvimento e aplicação da inteligência artificial. Espera-se que esta análise contribua para um debate mais informado e crítico incentivando a adoção de tecnologias e metodologias que minimizem o impacto ambiental e promovam a eficiência energética, garantindo um futuro mais sustentável e sem grandes comprometimentos para a tecnologia de inteligência artificial.

1 A INTELIGÊNCIA ARTIFICIAL NA ATUALIDADE

1.1 NOÇÕES GERAIS

A Inteligência Artificial tem se destacado como uma das áreas mais influentes e transformadoras da ciência da computação na atualidade. Esta tecnologia busca replicar a inteligência humana em máquinas, permitindo-lhes realizar tarefas como reconhecimento de fala, tomada de decisão e tradução de idiomas. Entre os subcampos da IA, destacam-se o aprendizado de máquina, a IA generativa, os Grandes Modelos de Linguagem (*Large Language Models* - LLM).

O Aprendizado de Máquina (*Machine Learning* - ML) é uma subárea crucial da IA, concentra-se no desenvolvimento de algoritmos que permitem aos computadores aprender e a partir de dados e fazer previsões ou decisões baseadas nesses dados. Mitchell (1997) define que "um programa de computador aprende com a experiência E em relação a alguma classe de tarefas T e medida de desempenho P, se seu desempenho em tarefas T, conforme medido por P, melhora com a experiência E". Este conceito é fundamental para a evolução de sistemas que podem melhorar seu desempenho com o tempo.

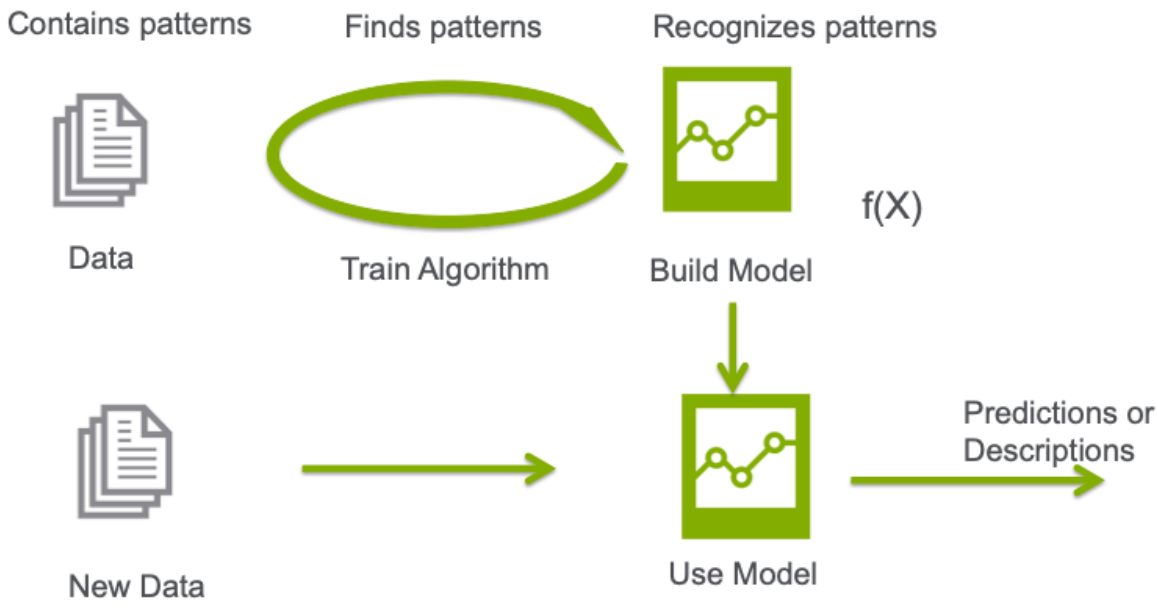

Figura 1: *What Is Machine Learning and How Does It Work?* Fonte: Nvidia *glossary*

O aprendizado federado é uma abordagem do ML onde os modelos são treinados em dados que permanecem localizados em dispositivos individuais, como *smartphones* e outros dispositivos pessoais. Em vez de centralizar os dados em um servidor, o modelo é treinado localmente em cada dispositivo e apenas os parâmetros atualizados do modelo (e não os dados brutos) são compartilhados com um servidor central. Esse servidor então agrega essas atualizações para melhorar o modelo global. Esta técnica é valiosa para preservar a privacidade dos dados dos usuários, reduzir a latência e a largura de banda necessária para enviar grandes volumes de dados, e é particularmente útil em aplicações onde a privacidade e a segurança dos dados são essenciais.

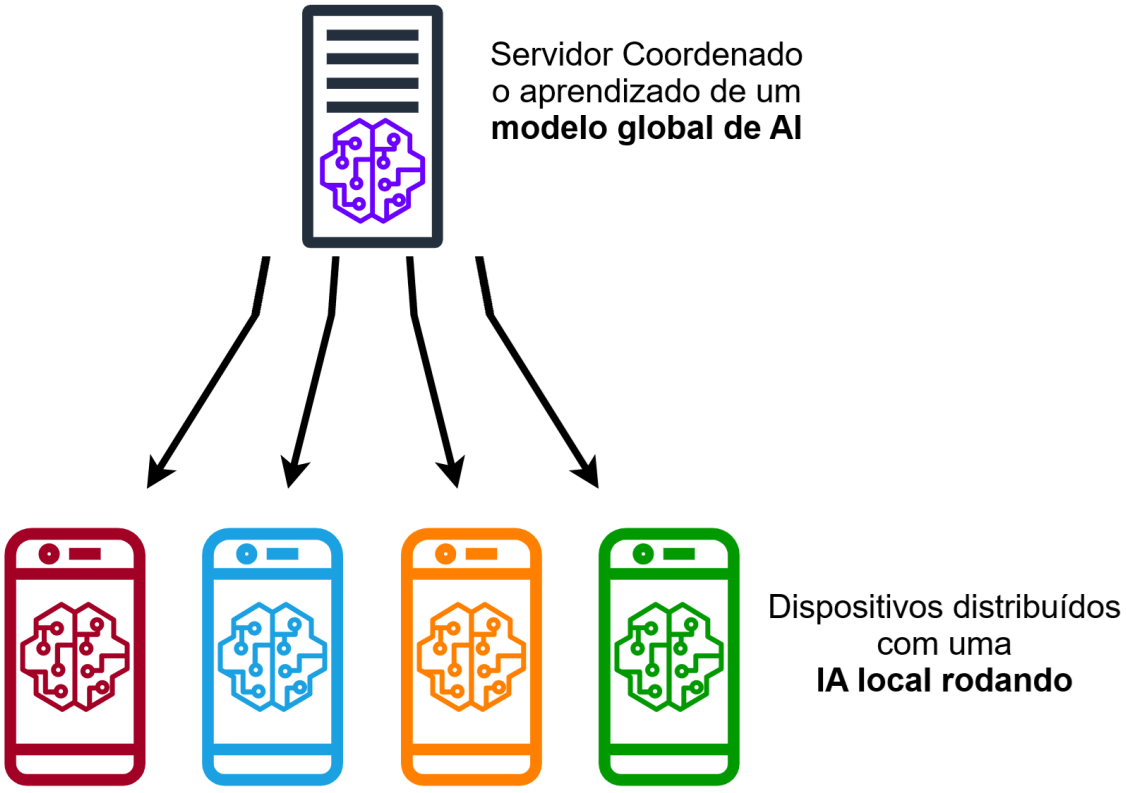

Figura 2: Aprendizado federado Fonte: Elaboração Própria

O aprendizado distribuído envolve a divisão do processo de treinamento de um modelo de *machine learning* em várias partes que são executadas em paralelo em diferentes máquinas ou nós de uma rede. Esses nós podem ser computadores em um *data center* ou em uma infraestrutura de nuvem. O aprendizado distribuído é

utilizado para acelerar o processo de treinamento e para lidar com conjuntos de dados e modelos que são grandes demais para serem processados por um único sistema. Os dados são distribuídos entre os nós, e cada nó processa uma parte do treinamento, enviando os resultados intermediários para um servidor central que coordena e consolida os resultados.

Os modelos de Aprendizado profundo (*Deep Learning)* são uma ramificação sofisticada do Aprendizado de Máquina, que utilizam grandes volumes de dados para realizar tarefas complexas, como reconhecimento de imagens, processamento de linguagem natural e previsão de séries temporais. Estes modelos são treinados para detectar padrões em dados e fazer previsões ou gerar saídas a partir de novas entradas.

Os LLMs são uma vertente avançada do Aprendizado de Máquina, eles utilizam vastas quantidades de dados textuais para gerar e compreender a linguagem natural de maneira eficaz, sendo um tipo específico de IA Generativa. A IA Generativa, por outro lado, é um campo que abrange uma ampla gama de sistemas dedicados à produção de conteúdo novo e inédito, como texto, imagens, música e código. Esses aplicativos de IA generativa são construídos com base em LLMs e modelos de fundação.

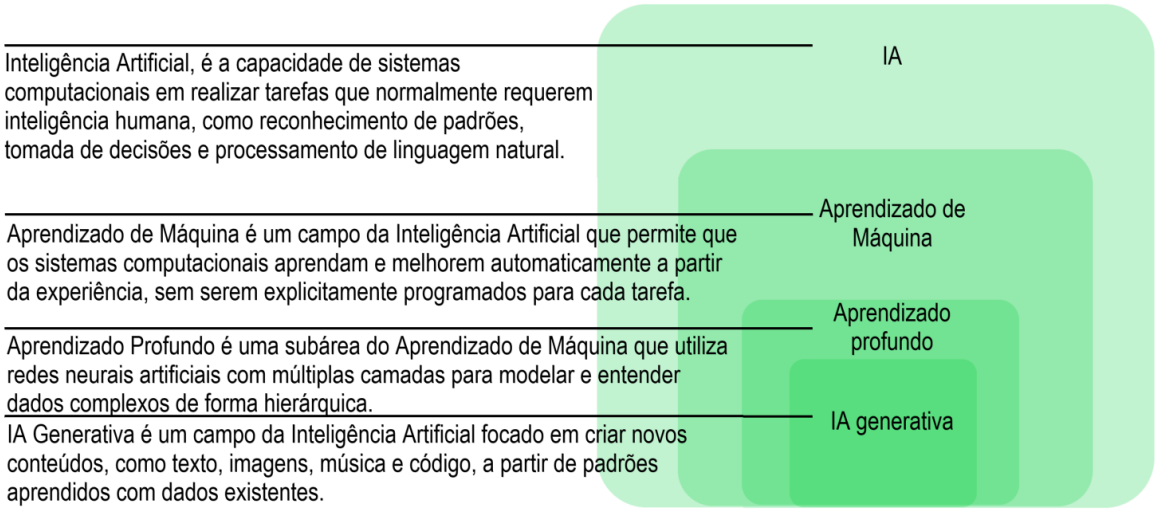

Figura 3: Campos da IA Fonte: Elaboração Própria

A aplicação dessas tecnologias é ampla e variada, abrangendo setores como saúde, finanças, transporte e entretenimento. Elas proporcionam automação de tarefas repetitivas, aprimoramento na tomada de decisões e criação de novas oportunidades de negócios, demonstrando uma influência positiva indiscutível na sociedade, o treinamento de modelos de IA modernos, especialmente os LLMs, exige grandes quantidades de energia, o que suscita preocupações sobre sustentabilidade energética e impactos ambientais.

1.2 IMPACTOS AMBIENTAIS DO TREINAMENTO DE IA

O avanço O treinamento de modelos de Inteligência Artificial (IA) tem demonstrado ser um processo intensivo em termos de recursos computacionais, resultando em significativos impactos ambientais. Entre esses impactos, destacam-se a emissão de gases de efeito estufa e o uso de água, que são aspectos críticos a serem considerados para uma abordagem sustentável na área de tecnologia.

A emissão de Gases de Efeito Estufa (GEE) é um dos principais impactos ambientais decorrentes do treinamento de modelos de IA. Segundo Strubell, Ganesh e McCallum (2019), o treinamento de grandes modelos de IA consome vastas quantidades de energia elétrica, grande parte dela proveniente de fontes não renováveis, como carvão e gás natural. Este consumo energético resulta na emissão de dióxido de carbono (CO_2) e outros gases de efeito estufa, contribuindo significativamente para as mudanças climáticas.

Estudos indicam que o treinamento das IAs, especialmente aqueles de grande escala como os LLMs, pode gerar emissões comparáveis às de um carro em toda a sua vida útil, incluindo a fabricação do veículo e o combustível consumido (Strubell et al., 2019). Patterson et al. (2021) destacam que a pegada de carbono associada ao treinamento desses modelos pode ser mitigada através da utilização de fontes de energia renováveis e da otimização dos algoritmos para reduzir o consumo de energia.

O uso de água é outro impacto ambiental relevante no contexto do treinamento de IA. A água é essencial para manter a temperatura dos servidores em níveis operacionais seguros, evitando o superaquecimento e garantindo a eficiência dos sistemas. No entanto, a dependência desse recurso natural para fins de

resfriamento apresenta um desafio significativo. Segundo Patterson et al. (2021), a adoção de tecnologias de resfriamento mais eficientes e a reutilização de água em circuitos fechados são estratégias promissoras para reduzir o impacto ambiental associado ao uso de água em banco de dados.

A mitigação das emissões de gases de efeito estufa e a gestão eficiente do uso de água são essenciais para minimizar o impacto ambiental da IA. A pesquisa e o desenvolvimento de novas tecnologias é um passo crucial para garantir um futuro mais sustentável de integração da tecnologia IA.

1.2.1 CONSUMO ENERGÉTICO EM MODELOS DE IA

Alguns estudos alertam sobre uma iminente explosão no consumo de energia pelos *data centers* de IA. De acordo com um estudo recente feito pela Dra. Luccioni, sistemas de IA generativa podem consumir até 33 vezes mais energia do que máquinas projetadas para executar tarefas específicas (LUCCIONI; JERNITE; STRUBELL, 2023).

Apesar dessas preocupações, existe uma vertente na área de tecnologia que sugere que o problema pode ser mitigado pela evolução contínua das tecnologias, que contribuirá para manter o consumo de energia em níveis aceitáveis. Contudo, diversos fatores influenciam a demanda de energia elétrica necessária para o processamento de inteligência artificial. Contudo, fabricantes de chips como a NVIDIA estão desenvolvendo componentes capazes de processar tarefas de IA diretamente nos PCs dos usuários, conforme destacado pelos analistas Nadkarni e Rutten (2023). Essa abordagem reduz a necessidade de processamento em data centers na nuvem, mas, logicamente, resulta em um aumento no consumo energético doméstico.

Portanto, para enfrentar o desafio de garantir a eficácia energética da IA, é essencial adotar técnicas de otimização energética amplas. A exemplo de soluções como o uso de algoritmos mais eficientes, que assim como demonstrado pelo pesquisado Mohammad Ali Khoshkholghi (2017), podem reduzir significativamente o consumo de energia. Além de manter a autonomia para que empresas possam continuar investindo recursos próprio no desenvolvimento de *hardware* e que os usuários possam usufruir dos mesmos.

Em conclusão, o consumo energético em modelos de IA é um desafio significativo que requer uma abordagem multifacetada. A conscientização e a educação sobre esses desafios são fundamentais para mobilizar esforços em direção a um futuro mais sustentável garantindo que o desenvolvimento da IA continue de forma sustentável e ativa.

1.2.2 DESAFIOS ENERGÉTICOS EM IA

O avanço da tecnologia de IA tem sido notável nas últimas décadas, trazendo consigo inúmeros benefícios para diversas áreas. No entanto, a crescente complexidade e o uso extensivo de modelos de IA, especialmente os LLMs, têm gerado preocupações significativas quanto ao consumo de energia.

A demanda por IA tem se mantido em constante crescimento, impulsionada pela capacidade dos modelos de processar e analisar grandes volumes de dados, oferecendo *insights* valiosos e automatizando tarefas complexas. No entanto, o treinamento de modelos de IA é um processo intensivo em termos de recursos, exigindo grandes quantidades de energia para alimentar tanto o processamento quanto a refrigeração dos *data centers*. De acordo com um estudo recente, o consumo energético dos *data centers* é responsável por aproximadamente 2% do consumo global de energia, e essa porcentagem tende a aumentar com a popularização da IA (Smith et al., 2022).

Para enfrentar esses desafios, é essencial adotar técnicas de otimização energética. Algoritmos mais eficientes podem reduzir significativamente o consumo de energia, enquanto a implementação de *data centers* sustentáveis, que utilizam fontes de energia renovável e tecnologias avançadas de refrigeração, pode minimizar o impacto ambiental. Além disso, inovações como o uso de discos ópticos 3D em escala nanométrica com capacidade de *petabit* têm o potencial de transformar o armazenamento de dados, oferecendo soluções mais eficientes e sustentáveis (Smith et al., 2022).

A otimização energética no treinamento de modelos de IA é crucial para tornar os sistemas mais eficientes e sustentáveis. É essencial que o uso dessas tecnologias seja guiado por princípios éticos, levando em consideração as responsabilidades ambientais e os valores inerentes ao desenvolvimento

sustentável. Essa abordagem é fundamental para maximizar os benefícios da IA, ao mesmo tempo em que se evitam problemas ecológicos e o consumo excessivo de energia. A adoção de práticas de otimização energética, juntamente com políticas regulatórias e tecnologias inovadoras, é essencial para assegurar que o avanço da IA ocorra de maneira consciente e sustentável. Isso envolve equilibrar a necessidade de alto desempenho com a responsabilidade ambiental, garantindo um futuro em que o desenvolvimento tecnológico e a sustentabilidade caminhem lado a lado. Este trabalho visa contribuir para esse debate, oferecendo *insights* e recomendações para a comunidade acadêmica e profissionais da área.

2 - ESTRATÉGIAS DE OTIMIZAÇÃO ENERGÉTICA

O treinamento de modelos de IA demanda uma quantidade significativa de recursos computacionais e energia elétrica, levando a impactos ambientais consideráveis. Logo, a implementação de estratégias de otimização energética é crucial para mitigar esses impactos e promover a sustentabilidade no desenvolvimento de tecnologias de IA.

Estudos como o de Pattnaik et al. (2021) têm destacado a importância de otimizar tanto o *hardware* quanto os algoritmos utilizados no treinamento de IA para alcançar uma eficiência energética superior. A pesquisa aponta que a combinação de *hardwares* especializados e algoritmos otimizados pode resultar em uma redução significativa no consumo de energia.

Dessa forma, para promover o desenvolvimento sustentável no campo da IA, é essencial investir em tecnologias de *hardware* e *software* eficientes, além de adotar práticas que visem a redução do consumo energético durante o treinamento de modelos. O avanço contínuo nessas áreas pode levar a um futuro em que a inteligência artificial seja utilizada de maneira sustentável a nível global.

2.1 HARDWARE EFICIENTE

Pesquisas recentes têm demonstrado avanços significativos na criação de

hardwares dedicados e otimizados para operações de IA. Um exemplo notável são as unidades de processamento gráfico (GPUs) e os aceleradores de IA, como as unidades de processamento tensorial (TPUs) desenvolvidas pelo Google. Esses dispositivos são projetados para realizar operações massivamente paralelas, o que permite um processamento mais rápido e eficiente de grandes volumes de dados, reduzindo assim o consumo energético (Jouppi et al., 2017; Spiridonov, 2021).

Além disso, essas inovações tecnológicas têm facilitado a implementação de modelos de aprendizado profundo mais complexos, como redes neurais convolucionais (CNNs) e redes neurais recorrentes (RNNs), que são fundamentais para aplicações em visão computacional e processamento de linguagem natural, respectivamente (LeCun et al., 2015; Goodfellow et al., 2016).

Outro aspecto importante é a integração de técnicas de otimização de *hardware*, como a quantização e a poda, que ajudam a reduzir a complexidade dos modelos sem comprometer significativamente a precisão. Tais técnicas são reconhecidas na otimização de modelos de IA desempenhando um papel significativo na mitigação dos impactos ambientais dessas tecnologias. Segundo Han, Mao e Dally (2016), a quantização, que consiste na redução da precisão dos parâmetros dos modelos, diminui a demanda por recursos computacionais sem afetar drasticamente a precisão, culminando em um menor consumo de energia. De forma complementar, a poda, ao remove neurônios ou conexões que não são essenciais, simplifica o modelo e aumenta sua eficiência operacional (Gale et al., 2019). A aplicação combinada dessas estratégias não só eleva a eficiência computacional, como também diminui a pegada de carbono associada aos processos de treinamento e inferência em IA, favorecendo práticas mais sustentáveis e conscientes no campo da tecnologia (Strubell, Ganesh e McCallum, 2019).

2.2 ALGORITMOS EFICIENTES

Diante do exposto, vale ressaltar que esses avanços não se limitam apenas ao *hardware* em si, mas também incluem melhorias nos algoritmos de treinamento e inferência. Métodos como o aprendizado federado e o aprendizado distribuído permitem o treinamento de modelos em dispositivos periféricos, minimizando a

necessidade de transferência de grandes volumes de dados para servidores centrais e aumentando a privacidade dos dados (Konečný et al., 2016; McMahan et al., 2017).

O aprendizado federado e o aprendizado distribuído são duas abordagens complementares que podem ser aplicadas para promover o desenvolvimento sustentável através da IA. O aprendizado federado, ao manter os dados localmente nos dispositivos dos usuários, reduz a necessidade de grandes infraestruturas de armazenamento centralizado e a transferência de dados em massa, o que economiza energia e reduz a pegada de carbono. Além disso, ao proteger a privacidade dos dados, promove a confiança dos usuários em sistemas de IA, incentivando uma adoção mais ampla e responsável da tecnologia.

Por outro lado, o aprendizado distribuído permite que grandes modelos de IA sejam treinados de forma mais eficiente, utilizando recursos de computação distribuídos. Podendo assim, levar a avanços mais rápidos em soluções de IA que por sua vez podem abordar questões ambientais e sociais, como monitoramento climático, otimização de energia, e gerenciamento de recursos naturais. Ao dividir a carga de trabalho entre várias máquinas, o aprendizado distribuído pode também utilizar fontes de energia renovável mais eficientemente, contribuindo para práticas mais sustentáveis.

Finalmente, o impacto desses avanços tecnológicos é evidente em diversas indústrias. Sistemas de IA estão sendo utilizados para a detecção precoce de doenças e a personalização de tratamentos, enquanto no setor automobilístico, tecnologias de condução autônoma estão cada vez mais aprimoradas graças aos algoritmos de aprendizado profundo e à computação de alto desempenho (Esteva et al., 2017; Bojarski et al., 2016). À medida que se continua a explorar e desenvolver essas tecnologias, o potencial para transformar diversos aspectos da sociedade e da economia permanece vasto e promissor.

2.3 ADOÇÃO DE COMPUTAÇÃO EM NUVEM

A computação em nuvem promove a concentração de recursos em *data centers* projetados para otimização do consumo de energia. Os fornecedores de *cloud computing* estão empenhados em aprimorar as tecnologias de eficiência

energética e adotar fontes de energia renováveis, o que tende a diminuir o consumo energético necessário para treinar e executar modelos de inteligência artificial, quando comparado com instalações locais que possuem menor otimização. Farjana et al. (2023) destacam que esta abordagem não só melhora a eficiência energética, mas também contribui significativamente para a redução de *hardware* necessário e, por consequência, diminui o consumo de materiais e a geração de resíduos eletrônicos.

Além disso, a computação em nuvem elimina a necessidade das empresas de manterem grandes infraestruturas de *hardware* próprio, facilitando a virtualização e o uso compartilhado de recursos, o que eleva a eficiência geral dos sistemas. Rajkumar Buyya e Sukhpal Singh Gill (2018) apontam que essa estrutura flexível também possibilita a escalabilidade dos recursos computacionais de acordo com a demanda, permitindo que as organizações ajustem a capacidade computacional de forma dinâmica para atender precisamente as necessidades operacionais, o que previne o desperdício de energia e de outros recursos.

Em resumo, a adoção da computação em nuvem pode tornar a inteligência artificial mais sustentável ao maximizar a eficiência energética, diminuir a demanda por *hardware*, permitir a escalabilidade sob demanda, otimizar o uso de recursos, promover o emprego de energia renovável, fomentar a inovação, e centralizar a manutenção e atualização de infraestruturas (Tomarchio et al., 2020). Além disso, a pesquisa de Katal, Dahiya, e Choudhury (2023) reforça que a eficiência energética em *data centers* de computação em nuvem é um campo em constante evolução, com inúmeras inovações em tecnologias de *software* sendo continuamente desenvolvidas para melhorar ainda mais essa eficiência.

3 - PERSPECTIVAS FUTURAS PARA A EFICIENCIA ENERGETICA

3.1 TECNOLOGIAS PROMISSORAS

Uma das principais promessas para enfrentar o consumo energético das IAs é a melhoria da eficiência dos *data centers*. Tecnologias emergentes, como o armazenamento óptico em escala nanométrica, prometem revolucionar a eficiência energética desses centros de dados. De acordo com Miao Zhao et al. (2024), essa tecnologia permite aumentar a capacidade de armazenamento de discos de

memória óptica para o nível de *petabit* (125 *terabytes*) por meio de um sistema de armazenamento tridimensional. Este novo disco utiliza 100 camadas distintas para armazenar dados, contrastando com discos tradicionais que utilizam apenas uma única camada.

Esses sistema se baseia em fenômenos foto físicos relacionados às moléculas fluoróforas, que apresentam baixa emissão de luz em sistemas diluídos, mas que, em estados agregados ou restritos, exibem um aumento significativo da fluorescência devido à inibição das rotações e vibrações intramoleculares, conforme descrito por Pazini et al. (2020).

Para alcançar a leitura em nanoescala, foi necessário primeiro desenvolver um disco com um sistema de armazenamento tridimensional em nanoescala, o que foi é possível graças a filmes de foto resistência cobertos por um corante de emissão induzida por agregação que são estimulados por feixes de laser femtosegundo, o que resulta em pontos de gravação com tamanho de super-resolução. Esses pontos de gravação são significativamente menores do que os encontrados em DVDs e Blu-Rays, permitindo armazenar um volume maior de conteúdo. Todo esse avanço foi possível graças à emissão induzida por agregação (AIE), utilizando luminógenos com características de emissão induzida por agregação (AIEgens), como o hexaphenylsilole (HPS) e o ácido dietilenotriaminopentacético (DTPA), para gerar uma exposição multifotônica no foto iniciador isopropil tioxantona (ITX). Esse processo cria um material denominado AIE-DDPR, de base química HPS-ITX-DTPA, que, ao ser efetivamente inserido sobre o disco, funciona como uma camada de filme transparente altamente uniforme. Esta camada permite que cientistas utilizem lasers em escala de nanopartículas para realizar gravações de informações com alta precisão (Zhao et al., 2024).

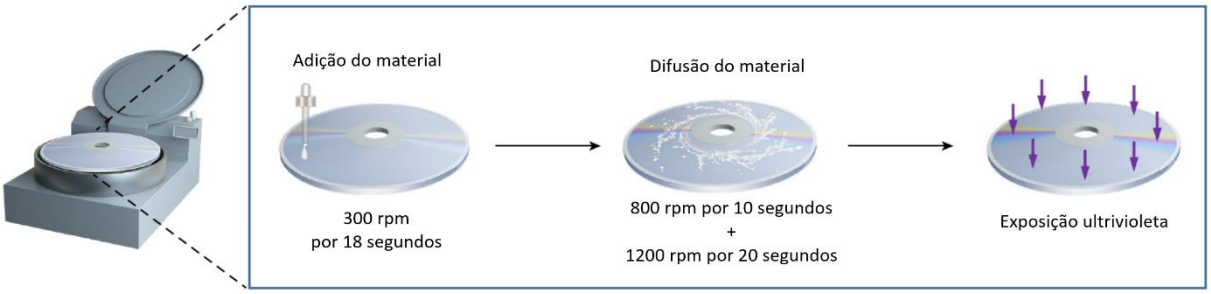

Figura 3: *Principle* of AIE-DDPR *disk* Fonte: Zhao et al., 2024

Em resumo, essa inovação tem o potencial de possibilitar a construção de *data centers* com capacidade de *exabit* (Eb) em espaços significativamente menores, como uma sala, em vez de áreas comparáveis a estádios, reduzindo drasticamente a pegada de carbono e os custos operacionais dos *data centers* (Zhao et al., 2024). Em resumo, a combinação de avanços tecnológicos em armazenamento de dados oferece uma perspectiva proeminente para enfrentar os desafios energéticos associados ao treinamento de modelos de IA.

3.2 LEIS E ACORDOS

O treinamento de modelos de IA está no coração de muitas inovações tecnológicas, mas também envolve desafios energéticos consideráveis. O foco crescente na eficiência energética e nas práticas sustentáveis, especialmente no contexto de regulamentações como o GDPR (*General Data Protection Regulation*) da União Europeia e a LGPD (Lei Geral de Proteção de Dados) do Brasil, bem como iniciativas como o *Climate Neutral Data Centre Pact*, influencia diretamente as perspectivas futuras para a indústria de IA.

O GDPR e a LGPD não tratam diretamente da eficiência energética. No entanto, ao exigir que as organizações sejam transparentes quanto ao uso e armazenamento de dados, essas leis encorajam indiretamente a adoção de infraestruturas mais eficientes. Isso pode incluir a otimização de *data centers* para o consumo reduzido de energia, o que é essencial quando se considera o grande volume de dados necessários para treinar modelos de IA. A transparência e a conformidade com essas regulamentações podem levar as empresas a buscar tecnologias de armazenamento e processamento de dados que sejam não apenas seguras, mas também energeticamente eficientes.

O *Climate Neutral Data Centre Pact* é uma iniciativa significativa que promove diretrizes claras para a sustentabilidade em *data centers*, com metas de alcançar neutralidade climática até 2030. Esse pacto pressiona os provedores de serviços a inovar em soluções de eficiência energética e a adotar fontes de energia renováveis. Para o treinamento de modelos de IA, isso significa que as empresas que fornecem recursos computacionais precisarão adaptar suas operações para serem menos

dependentes de combustíveis fósseis e mais focadas em alternativas sustentáveis. A transição para energias renováveis e a melhoria da eficiência energética dos data centers são passos cruciais para mitigar os impactos ambientais associados ao treinamento de IA.

Com o passar do tempo, é provável que as restrições e necessidades energéticas no treinamento de IA se tornem um campo cada vez mais central. Portanto, o campo de IA enfrenta um desafio duplo: continuar a inovar enquanto se adapta a práticas mais sustentáveis. As empresas de tecnologia terão que investir em pesquisa e desenvolvimento de algoritmos e tecnologias que não apenas sejam eficientes em termos de desempenho, mas que também minimizem o consumo de energia durante o treinamento e a operação.

CONCLUSÃO

Portanto, a partir do que foi exposto anteriormente, torna-se evidente que o desenvolvimento e a aplicação de Inteligência Artificial acarretam desafios ambientais significativos, sobretudo relacionados ao seu extenso consumo energético. No entanto, com a introdução de estratégias inovadoras como *hardware* eficiente, algoritmos otimizados e a adesão à computação em nuvem, se observa que é possível mitigar substancialmente esses impactos. As perspectivas futuras são igualmente promissoras, considerando as tecnologias emergentes que prometem a integração sustentável dessa tecnologia na sociedade. Tecnologias como discos ópticos 3D em escala nanométrica e fontes de energia renováveis podem revolucionar a forma como os *data centers* operam. Além disso, é essencial que as regulamentações continuem a evoluir, e que sejam implementadas de maneira consistente, garantindo que as soluções de IA não apenas minimizem seu impacto ambiental, mas também promovam práticas de desenvolvimento sustentável. Neste contexto, a colaboração entre o setor público, a academia e a indústria é fundamental para criar um ecossistema onde a inovação em IA possa prosperar sem comprometer o meio ambiente. Projetos de pesquisa colaborativa, incentivos para práticas sustentáveis e a implementação de políticas que favoreçam a eficiência energética são passos importantes nessa direção. Por fim, espera-se que este trabalho contribua para um debate mais informado e crítico, promovendo a sustentabilidade como um pilar central no desenvolvimento contínuo da IA garantindo que seus benefícios possam ser aproveitados de maneira responsável e consciente. Dessa forma, pode-se assegurar um futuro em que a inovação tecnológica e a preservação ambiental caminhem lado a lado, beneficiando tanto a sociedade quanto o planeta.

Trabalhos Futuros

A fim de continuar avançando na mitigação dos impactos ambientais da IA, futuros trabalhos podem se concentrar abrangir;

Incorporação de tecnologias emergentes, como discos ópticos 3D em escala

nanométrica, para otimizar ainda mais o armazenamento de dados.

Desenvolvimento e a aplicação de algoritmos mais eficientes em termos de energia.

Implementação de estratégias de aprendizado federado e distribuído para reduzir a necessidade de grandes *data centers*.

Investigação sobre a utilização de fontes de energia renováveis específicas para data centers e infraestruturas de IA.

Análise e proposta de novas políticas e regulamentações que incentivem práticas sustentáveis no desenvolvimento e aplicação de IA.

Avaliação dos impactos sociais e econômicos das tecnologias de IA sustentáveis, promovendo um equilíbrio entre inovação tecnológica e responsabilidade ambiental.

REFERENCIAS

JONES, A. Energy consumption in data centers: a review. Journal of Sustainable Computing, v. 15, n. 2, p. 123-135, 2021.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997.

SMITH, B. et al. Advances in data storage: the future of 3D optical disks. IEEE Transactions on Nanotechnology, v. 21, n. 3, p. 45-52, 2022.

LUCCIONI, Alexandra Sasha; JERNITE, Yacine; STRUBELL, Emma. Power hungry processing: Watts driving the cost of ai deployment? 2023.

NADKARNI, Ashish; RUTTEN, Peter. Demystifying Generative AI. Analyst Brief. Sponsored by: NVIDIA, 2023.

SMITH, John. Energy Consumption in Data Centers. Analyst Brief. New York: Tech Insights, 2022.

LUCCIONI, Sasha; JERNITE, Y.; STRUBELL, E. Energy Consumption in AI: A Critical Review. Journal of AI Research, 2023.

NADKARNI, A.; RUTTEN, J. The Future of AI Processing: From Cloud to Local Devices. Tech Insights, 2023.

KHOSHKHOLGHI, M. A.; DERAHMAN, M. N.; ABDULLAH, A.; Energy-Efficient Algorithms for Dynamic Virtual Machine Consolidation in Cloud Data Centers. Analyst Brief. 2017.

GALE, R. C. et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. Implementation Science, v. 14, n. 11, 2019.

HAN, S.; MAO, H.; DALLY, W. J. Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding. In: 4th International Conference on Learning Representations, ICLR 2016.

STRUBELL, E.; GANESH, A.; MCCALLUM, A. Energy and Policy Considerations for Deep Learning in NLP. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2019.

SURDEN, Harry. The Ethics of Artificial Intelligence in Law: Basic Questions. 1. ed. Oxford: Oxford University Press, 2019.

SPIRIDONOV, Alexander. New Cloud TPU VMs make training your ML models on TPUs easier than ever. 2021.

PAZINI, A. et al. Designing highly luminescent aryloxy-benzothiadiazole derivatives with aggregation-induced enhanced emission. Dye. Pigment., v. 178, 2020

ZHAO, Miao; WEN, Jing; HU, Qiao; WEI, Xunbin; ZHONG, Yu-Wu; RUAN, Hao; GU, Min. A 3D nanoscale optical disk memory with petabit capacity. *Nature Nanotechnology*, [S.l.], v. 19, n. 3, p. 345-350, 2024