Aprendizagem Profunda e Inteligência Artificial Verde: Caminhos para um Futuro mais Sustentável

Vívian R. G. Ferraro[1], Gabriel Gullo[1], Daniel da Silva Costa[2], Pedro Nuno de S. Moura[2]

[1]Escola de Informática Aplicada (EIA)
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

[2]Programa de Pós-Graduação em Informática (PPGI)
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

{viferraro, gabriel.gullo, daniel.scosta}@edu.unirio.br, pedro.moura@uniriotec.br

Abstract. *In the last decade, there have been significant advances in the results achieved by Deep Learning models and their widespread adoption in academia and industry. Although these models have the potential to assist in the management of natural resources and environmental issues, they typically require a great deal of computing power, resulting in higher energy costs and also large carbon footprint numbers. This work seeks to highlight and discuss the energy costs involved in using neural network models, experimentally comparing some architectures in terms of performance, energy efficiency, and computational cost. The results obtained reinforce that it is possible to build models that consume less energy and have performance compatible with more expensive ones, contributing to a more sustainable approach.*

Resumo. *Na última década, houve avanços significativos nos resultados alcançados por modelos de Aprendizagem Profunda e uma ampla adoção desses na academia e na indústria. Embora esses modelos tenham potencial para auxiliar na gestão de recursos naturais e em questões ambientais, eles tipicamente demandam grande poder computacional, resultando em maiores gastos energéticos e também em grandes números de pegada de carbono. Este trabalho busca evidenciar e discutir o gasto energético envolvido no uso de modelos de redes neurais, comparando experimentalmente algumas arquiteturas em relação ao desempenho, à eficiência energética e ao custo computacional. Os resultados obtidos reforçam que é possível construir modelos que consumam menos energia e que tenham desempenho compatível com aqueles mais dispendiosos, contribuindo para uma abordagem mais sustentável.*

1. Introdução

A Inteligência Artificial (IA) tem se tornado uma ferramenta cada vez mais importante em diversos setores da sociedade, incluindo saúde, educação, transporte e segurança [Kelleher 2019]. Em particular, os modelos baseados em Aprendizagem Profunda (AP) têm demonstrado melhores resultados em comparação a outras abordagens de IA e parecem ser a principal escolha de abordagem nos dias atuais.

Uma das áreas de aplicação mais importantes atualmente, considerando-se os Objetivos de Desenvolvimento Sustentável[1] indicados pela Organização das Nações Unidas,

[1]https://www.unicef.org/brazil/objetivos-de-desenvolvimento-sustentavel

é o meio ambiente. Algumas questões ambientais que podem se beneficiar dos modelos de AP são, por exemplo: (i) o uso consciente dos recursos naturais; (ii) a preservação do meio ambiente; e (iii) o combate à mudança climática global. Em todos esses casos, os modelos de AP, por serem mais robustos e com reconhecido desempenho [Kelleher 2019], podem ajudar na obtenção de modelos mais fiéis sobre a dinâmica dos fenômenos ambientais e a utilização dos recursos naturais.

Modelos de AP são capazes de realizar tarefas complexas, mas tipicamente requerem um grande consumo de recursos computacionais [Wolff Anthony et al. 2020]. O treinamento desses modelos pode consumir grandes quantidades de energia devido ao grande volume de computação necessária. Desde 2012, os modelos de AP têm registrado um desempenho crescentemente superior, mas o custo de treinamento aumentou por um fator de 300.000 nesse período, o que não é sustentável [Lenherr et al. 2021]. Ademais, o próprio descarte de equipamentos utilizados para treinar ou executar modelos de AP também pode gerar um impacto ambiental significativo.

O impacto ambiental do desenvolvimento e uso de modelos de AP é uma questão importante que precisa ser considerada. As pesquisas e aplicações na área de IA estão em uma fase de bastante efervescência, e alguns trabalhos recentes apontam para a necessidade de adotar medidas para minimizar esse impacto [Schwartz et al. 2020, Wolff Anthony et al. 2020, Lenherr et al. 2021, Desislavov et al. 2023].

Em meio aos avanços da IA, surgiu o conceito de Inteligência Artificial Verde (IA Verde), uma abordagem responsável que busca harmonizar o progresso tecnológico com a sustentabilidade ambiental. Conforme introduzido por [Schwartz et al. 2020], a IA Verde se opõe à IA Vermelha, que prioriza o aumento da acurácia sem considerar os custos ambientais, econômicos e sociais. A IA Verde enfatiza a eficiência como um critério de avaliação essencial, promovendo práticas que reduzam o consumo de energia e as emissões de carbono. Isso inclui, por exemplo, a adoção de técnicas como a quantização, que otimiza o uso de recursos computacionais e energéticos, ao reduzir a precisão dos parâmetros do modelo e diminuir o tamanho do modelo, ou o compartilhamento de modelos pré-treinados, que evita a redundância no uso de recursos.

Este trabalho busca demonstrar a viabilidade de se obterem modelos de AP mais eficientes e com menor impacto ambiental. Para tal, foi conduzida uma comparação experimental de arquiteturas de redes neurais para visão computacional considerando métricas tradicionais de avaliação de desempenho associadas a métricas de eficiência energética e de custo computacional. Os resultados obtidos mostram que é possível construir modelos que consumam menos energia e que tenham menor custo computacional, mas que possuam uma performance similar àquela de modelos mais dispendiosos, corroborando, assim, a proposta da IA Verde.

O restante deste artigo está estruturado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados que abordam diferentes aspectos da eficiência energética e do custo computacional de modelos de AP; já a Seção 3 descreve os materiais e métodos empregados nos experimentos computacionais realizados; por sua vez, a Seção 4 descreve em detalhes os resultados obtidos e os discute; e, por fim, a Seção 5 apresenta as considerações finais e os trabalhos futuros.

2. Trabalhos Relacionados

A preocupação com o consumo energético por modelos de AP é uma questão bastante recente e que passou a atrair a atenção da academia e da indústria à medida que tais modelos foram crescendo em números de parâmetros, aumentando cada vez mais o seu desempenho, mas também a sua demanda energética.

Nesse sentido, o trabalho de [Schwartz et al. 2020] foi o primeiro a se dedicar a esta questão e a definir o conceito de "IA Verde". Uma proposta dos autores correspondeu a usar o número total de FLOPs como uma medida de eficiência, incentivando os pesquisadores a reportarem essa métrica. Os autores defendem que a IA Verde é mais inclusiva e sustentável do que a IA Vermelha, apontando diversas direções de pesquisa para avançar a eficiência na IA.

Após isso, alguns outros trabalhos buscaram se dedicar a entender melhor esta questão e a abordar diferentes aspectos da eficiência, do consumo energético e do custo computacional de modelos de AP. Em [Wolff Anthony et al. 2020], foi proposto o Carbontracker, uma ferramenta que monitora e prevê o consumo de energia e as emissões de carbono do treinamento de modelos de AP. A ferramenta visa conscientizar os pesquisadores sobre o impacto ambiental dos seus trabalhos, de modo a incentivá-los a adotar práticas mais sustentáveis. O artigo também avaliou a ferramenta em diferentes arquiteturas de redes neurais e conjuntos de dados para segmentação de imagens médicas e fornece recomendações para diminuir as emissões de carbono.

O estudo de [Lacoste et al. 2019], investigou o impacto ambiental do treinamento de redes neurais, mostrando que diversos fatores influenciam a emissão de carbono, como localização do servidor, fonte de energia, duração do treinamento e hardware utilizado. Para auxiliar na compreensão desse impacto, os autores apresentaram a Calculadora de Emissões de Aprendizagem de Máquina, ferramenta que estima as emissões de acordo com o tipo de energia e infraestrutura de computação utilizados. Os autores esperam que o estudo incentive discussões e ações para reduzir o impacto ambiental da pesquisa em IA.

Em [Dhar 2020], o autor discute o impacto da utilização da IA no meio ambiente. De um lado, a IA tem o potencial de contribuir para a luta contra a crise climática, com projetos de infraestrutura sustentável e previsões meteorológicas mais precisas. Por outro lado, o custo ambiental de operar e treinar modelos de IA é significativo, com um único modelo de linguagem podendo emitir tanto carbono quanto 125 voos transcontinentais, de modo que o autor adverte sobre a necessidade de transparência nas emissões de carbono dos modelos de IA, citando como exemplo a ferramenta desenvolvida por [Lacoste et al. 2019].

Por sua vez, o trabalho de [Douwes et al. 2021] abordou o problema do consumo de energia de modelos geradores de áudio baseados em AP. Os autores propuseram uma nova métrica baseada na fronteira de Pareto, que considera tanto a qualidade do som sintetizado quanto o consumo de energia do modelo. Os autores aplicaram essa métrica em vários modelos existentes e mostraram que essa pode alterar significativamente os resultados obtidos. Além disso, também realizaram experimentos com um modelo recentemente proposto chamado WaveFlow, e mediram o consumo de energia real de treinamento e inferência para cinco configurações alternativas. Eles usaram um espaço multiobjetivo para

representar as soluções ótimas na fronteira de Pareto e analisaram o compromisso entre qualidade da saída gerada e eficiência energética dos modelos.

Já o trabalho de [Desislavov et al. 2023] analisou o consumo de energia em modelos de visão computacional e processamento de linguagem natural (PLN), com um foco especial em seus custos de inferência e não apenas no treinamento. Os autores argumentam que, para um sistema em produção, o custo de inferência ultrapassa o de treino, por causa do fator multiplicativo oriundo dos inúmeros usos em produção. O estudo envolveu a coleta de dados sobre o desempenho de modelos da literatura, o número de parâmetros, o número de operações de ponto flutuante (FLOPs) associadas às inferências, de modo que os autores estimaram o consumo de energia por inferência usando dados de hardware de placas gráficas (GPUs) da NVIDIA. Os resultados mostraram que, embora o consumo de energia dos modelos de ponta ainda cresça exponencialmente para um aumento sustentado no desempenho, o ritmo é mais lento do que o esperado, graças às melhorias algorítmicas e à especialização do hardware. No entanto, os autores terminam o artigo alertando para o efeito multiplicativo associado à penetração cada vez maior da IA na sociedade, que pode elevar o consumo de energia global.

O trabalho de [Lenherr et al. 2021] propôs novas métricas, que equilibram o consumo energético e o desempenho de um modelo, para medir a eficiência e a sustentabilidade da AP em plataformas de borda. Tais métricas são baseadas no gradiente de reconhecimento, no número de classes identificadas e na energia gasta por inferência ou treinamento. O artigo também comparou a eficiência da AP em relação às abordagens centralizada e federada, e analisa o ciclo de vida da AP, incluindo o número de vezes que os modelos são usados. Por fim, o artigo também realizou uma comparação entre a eficiência das inteligências artificial e natural, fornecendo perspectivas sobre a escalabilidade da AP.

No trabalho de [Strubell et al. 2019], foram analisados o custo financeiro e ambiental do treinamento de modelos de AP para PLN. Para tal, os autores estimaram o consumo de energia, as emissões de carbono e o custo da computação em nuvem referentes ao treinamento de vários modelos populares de PLN. Foram propostas recomendações para reduzir os custos e melhorar a equidade na pesquisa e na prática de PLN, tais como: relatar o tempo de treinamento e a sensibilidade aos hiperparâmetros, prover acesso equitativo aos recursos computacionais, assim como priorizar o desenvolvimento de modelos e hardware eficientes.

3. Materiais e Métodos

Neste trabalho, foi adotada uma abordagem experimental para avaliar e comparar arquiteturas de redes neurais voltadas para a área de visão computacional, em relação a métricas de desempenho, de eficiência energética e de custo computacional.

As métricas adotadas, a serem apresentadas na Seção 3.1, foram aplicadas durante as fases de treinamento e de inferência dos modelos, para a base de dados MNIST [LeCun et al. 1998]. Após a aplicação das métricas em cada modelo, os resultados obtidos foram analisados utilizando-se técnicas estatísticas para facilitar a comparação dos modelos sob as perspectivas de performance e de eficiência energética. No que segue, os materiais e métodos empregados serão explicados detalhadamente.

3.1. Métricas

As métricas selecionadas foram destinadas a avaliar não apenas o desempenho técnico dos modelos de AP, mas também a sua sustentabilidade. As métricas de avaliação de desempenho correspondem àquelas comumente adotadas na literatura. As métricas de eficiência energética e de custo computacional foram selecionadas a partir do levantamento bibliográfico conduzido.

3.1.1. Métricas de Avaliação de Desempenho

As métricas comumente adotadas para avaliação de um modelo que aborda um problema de classificação correspondem a: acurácia, precisão e revocação [Aggarwal 2023]. Assim sendo, a métrica de acurácia expressa a taxa de acerto de um modelo e é calculada a partir da divisão da quantidade de predições corretas do modelo pela quantidade de predições feitas. A precisão mede o quanto um modelo prediz corretamente exemplos de uma dada classe e é calculada dividindo-se a quantidade de predições feitas corretamente para uma classe pela quantidade de exemplos que o modelo classificou como sendo dessa classe. Já a revocação calcula o quanto um modelo identifica corretamente exemplos de uma dada classe em relação a todos os exemplos dessa classe na base de dados e é calculada dividindo-se a quantidade de predições feitas corretamente para uma classe pela quantidade de exemplos existentes dessa classe. Por fim, a medida-F consiste na média harmônica entre a precisão e a revocação.

3.1.2. Métricas de Eficiência Energética e de Custo Computacional

O entendimento da eficiência energética dos modelos de AP é fundamental para a compreensão do impacto ambiental de sistemas computacionais. Assim, as métricas dessa categoria utilizadas neste trabalho incluem a quantidade de energia utilizada pela GPU, que é uma medida direta do consumo elétrico dos componentes de processamento gráfico durante a execução de tarefas, e a pegada de carbono (*carbon footprint*), que estima o impacto ambiental desse consumo, convertendo-o em equivalente de dióxido de carbono (CO_2) liberado na atmosfera.

Mais especificamente, a medição do consumo de energia das GPUs foi feita por meio da Biblioteca de Gerenciamento NVIDIA[2] (*NVIDIA Management Library* - NVML) que fornece dados precisos e em tempo real de diversos atributos de componentes NVIDIA, entre os quais o consumo energético. A unidade de medida adotada pela NVML é o miliwatt (mW), o que permite um monitoramento detalhado e preciso da energia consumida. Tais funcionalidades contribuem para uma gestão energética mais eficiente e sustentável das GPUs, possibilitando um uso mais responsável da energia.

Já a pegada de carbono foi aferida pela ferramenta Carbontracker[3], proposta por [Wolff Anthony et al. 2020], que permite acompanhar as emissões de carbono de um modelo de AP, em gramas. Essa informação pode ser convertida em números interpretáveis de atividades do dia a dia, como, por exemplo, a distância equivalente percorrida

[2]https://developer.nvidia.com/nvidia-management-library-nvml

[3]https://github.com/lfwa/carbontracker

por um carro, ajudando a contextualizar o impacto ambiental.

Foram também empregadas algumas métricas que refletem o custo computacional associado a um modelo, a saber: (i) número de parâmetros, que é a quantidade de parâmetros treináveis de um modelo; (ii) FLOPs, que representa o montante de computação (número de operações de ponto flutuante) para realizar uma determinada tarefa (e.g., uma inferência); (iii) tempo de treinamento, que se refere à quantidade de tempo tomada no treinamento de um modelo; (iv) tempo de inferência, que expressa a quantidade de tempo tomada ao se usar o modelo para inferência.

3.2. Arquiteturas Adotadas

As arquiteturas escolhidas para a realização dos experimentos deste trabalho foram: LeNet-5, AlexNet, ResNet e MobileNet. Essas foram selecionadas por serem amplamente adotadas na literatura, em *benchmarks* conduzidos na área, assim como por possuírem diferentes níveis de complexidade. Tais arquiteturas serão brevemente descritas a seguir.

A LeNet-5, proposta por [LeCun et al. 1998], foi um modelo pioneiro de rede neural convolucional (CNN) para a tarefa de reconhecimento de dígitos manuscritos. Essa arquitetura é relativamente simples, composta por duas camadas convolucionais e duas camadas totalmente conectadas subsequentes. Apesar de sua simplicidade, a LeNet-5 é capaz de alcançar ótimo desempenho no reconhecimento de dígitos e em outras tarefas de visão computacional.

Já a AlexNet [Krizhevsky et al. 2012] correspondeu a um marco na pesquisa de IA que renovou o interesse pelas redes neurais após os resultados surpreendentes obtidos na competição ILSVRC de 2012. Essa arquitetura é uma CNN mais complexa do que a LeNet-5, sendo composta por cinco camadas convolucionais e três camadas totalmente conectadas, além de algumas camadas de *pooling*. Acima de tudo, a AlexNet realizou algumas escolhas de projeto que passaram a ser o padrão em arquiteturas posteriores [Aggarwal 2023], tais como o uso da função de ativação ReLU e de *data augmentation*. Mesmo tendo sido proposta há mais de uma década, essa rede é capaz de atingir resultados expressivos para diversas tarefas de visão computacional.

A ResNet foi introduzida por [He et al. 2015] e representou um avanço significativo na pesquisa de CNNs. A principal inovação trazida foi a introdução de *conexões residuais* que permitem que camadas estejam conectadas a outras camadas posteriores, mas que não sejam adjacentes. Tais conexões ajudam a contornar o problema do desaparecimento do gradiente, em que esses assumem valores pequenos à medida que são propagados para trás através das camadas. Assim, com esse mecanismo, é possível treinar redes neurais mais profundas eficientemente, tirando proveito da capacidade da sua capacidade de aprendizado maior. Este trabalho utilizou a ResNet-34 que possui 34 camadas convolucionais.

Finalmente, a MobileNet, proposta por [Howard et al. 2017], é uma rede neural projetada para aplicações móveis. A MobileNet é otimizada para alcançar um equilíbrio entre precisão e eficiência, o que a torna ideal para dispositivos com recursos limitados. Ao invés de usar convoluções tradicionais, que aplicam filtros em todos os canais de entrada, a MobileNet utiliza a *convolução em profundidade*, que aplica um filtro separado para cada canal de entrada, e a *convolução ponto a ponto* que combina os resultados da convolução em profundidade. Essas estratégias reduzem o número de parâmetros a o

custo computacional. Neste trabalho, utilizou-se a MobileNetV1, que é a versão mais simplificada da família MobileNet.

3.3. Base de Dados

A base de dados escolhida para os experimentos foi a MNIST, proposta por [LeCun et al. 1998], que contém dígitos manuscritos coletados de estudantes do ensino médio e de trabalhadores do censo americano. Essa base é composta por um total de 70.000 imagens, divididas em 60.000 para treinamento e 10.000 para teste. Cada imagem representa um dígito de 0 a 9, em tons de cinza e com resolução de 28x28 *pixels*. A despeito de ser relativamente pequena, essa base é comumente adotada na validação experimental de métodos propostos para visão computacional, por conter imagens diversas e com detalhes suficientemente complexos [Krohn and Beyleveld 2020].

3.4. Ambiente Computacional

O ambiente computacional utilizado nos experimentos foi o seguinte: processador 11th Gen Intel(R) Core(TM) i5-11400, 32 GB DDR4 de memória RAM, placa de vídeo NVIDIA GeForce RTX 3060 e sistema operacional Ubuntu 22.04 LTS. A implementação dos modelos foi feita na linguagem de programação Python na versão 3.11, usando-se a biblioteca PyTorch[4] na versão 2.2. Os códigos desenvolvidos nesta pesquisa, assim como o resultado dos experimentos, estão disponíveis publicamente em um repositório criado no GitHub[5].

3.5. Avaliação Experimental

Na avaliação experimental, houve a adoção de duas fases: treinamento e inferência. Na primeira fase, foram treinados 10 modelos de cada uma das arquiteturas selecionadas, indicadas na Seção 3.2, de modo que foi computada a função de perda de cada modelo em relação aos conjuntos de treinamento e de validação, medindo-se também a eficiência energética, a pegada de carbono e o tempo de treinamento. Para cada arquitetura, foi então escolhido o modelo treinado associado à menor média do valor da função de perda para o conjunto de validação, para utilização na segunda fase. Nessa fase, para cada arquitetura, o modelo escolhido foi utilizado para inferência em relação ao conjunto de testes, por 10 vezes, tomando-se o tempo e as métricas de avaliação de desempenho. Por fim, foi calculado o tempo médio de inferência considerando essas 10 execuções. Ademais, os seguintes valores de hiperparâmetros foram usados: tamanho de lote de 64, número de épocas de 20, taxa de aprendizado de $0,001$ e otimizador *Adam*.

4. Resultados Obtidos

Inicialmente, são exibidos na Tabela 1 os valores das métricas de avaliação de desempenho para cada uma das redes testadas, em que o maior valor de cada métrica está destacado em negrito. Nessa tabela, é possível ver que a rede ResNet-34 foi aquela que obteve a melhor performance em relação às métricas de acurácia, precisão, revocação e medida-F, que corresponderam a $0,993$. Note-se que, para todas as redes, os valores das quatro métricas são os mesmos por conta do arredondamento feito com três casas decimais.

[4]https://pytorch.org/blog/pytorch2-2/

[5]https://github.com/viferraro/WCAMA

Por sua vez, a Tabela 2 apresenta os valores das métricas de eficiência energética, em que os menores valores estão destacados em negrito. Para facilidade de visualização, os valores de consumo energético foram convertidos de miliwatts para watts. Verifica-se, assim, que a LeNet-5 gerou o menor valor de pegada de carbono, que foi de $0,706$ gramas. Já para o valor energético, a rede MobileNetV1 foi a que menos consumiu energia, com $45,241$ watts gastos. No cômputo geral, percebe-se uma clara diferença, no que concerne a essas métricas, entre os valores obtidos pelas redes LeNet-5 e MobileNetV1, e aqueles das redes AlexNet e ResNet-34.

A Tabela 3 exibe os valores das métricas de custo computacional para as redes testadas, em que os menores valores estão em negrito. A MobileNetV1 possui o menor número de parâmetros e de FLOPs, à medida que a LeNet-5 está associada às menores quantidades de tempo de treino e tempo de inferência.

Tabela 1. Valores das métricas de acurácia, precisão, revocação e medida-F obtidos pelas redes utilizadas.

Rede	Acurácia	Precisão	Revocação	Medida-F
LeNet-5	0,987	0,987	0,987	0.987
AlexNet	0,992	0,992	0,992	0,992
ResNet-34	**0,993**	**0,993**	**0,993**	**0,993**
MobileNetV1	0,989	0,989	0,989	0,989

Tabela 2. Valores de emissão de carbono, em gramas, e gasto energético, em watts, obtidos pelas redes adotadas.

Rede	Carbono (g)	Energia (W)
LeNet-5	**0,706**	47,071
AlexNet	3,374	87,225
ResNet-34	2,721	83,110
MobileNetV1	0,807	**45,241**

Tabela 3. Valores de quantidade de parâmetros, FLOPs, tempo de treino, em segundos, e tempo de inferência, em segundos, para as redes usadas.

Rede	# Parâmetros	FLOPs	Tempo treino (s)	Tempo inferência (s)
LeNet-5	609.354	11.401.984	**160,666**	**1,075**
AlexNet	23.271.114	190.336.256	485,357	1,802
ResNet-34	21.292.042	69.921.536	413,479	1,321
MobileNetV1	**23.317**	**1.158.096**	168,605	1,166

Adicionalmente, a fim de se demonstrar que há, de fato, uma diferença estatisticamente significativa entre os valores de pegada de carbono e consumo de energia gerados pelas redes LeNet-5 e MobileNetV1 (grupo 1, constituído pelas redes menos dispendiosas), e AlexNet e ResNet-34 (grupo 2, constituído pelas redes mais dispendiosas), foram realizados dois testes de hipótese do tipo teste t de Student, um para cada métrica. A hipótese nula foi de que não há diferença significativa para os valores de pegada de carbono (consumo de energia) gerados pelos grupos 1 e 2. Já a hipótese alternativa foi de

que há diferença significativa para os valores de pegada de carbono (consumo de energia) gerados pelos grupos 1 e 2.

Para a métrica de pegada de carbono, adotou-se um p-valor de 0.020. O valor t calculado foi de -6.93, logo se rejeitou a hipótese nula, indicando que há uma diferença estatisticamente significativa entre os valores de pegada de carbono do grupo 1 e do grupo 2.

Analogamente, para a métrica de consumo energético, adotou-se um p-valor de 0.003. Assim, o valor de t calculado foi de -17.32, rejeitando-se a hipótese nula e mostrando que há uma diferença estatisticamente significativa entre os valores de consumo de energia do grupo 1 e do grupo 2.

4.1. Discussão dos Resultados

Os resultados demonstram que as redes AlexNet e ResNet-34, por disporem de um número de parâmetros e FLOPs significativamente maiores do que a LeNet-5 e a MobileNetV1, conseguiram alcançar os maiores valores de acurácia, precisão e revocação, todos acima de $0,99$. Isso está de acordo com resultados disponíveis na literatura, em que modelos mais complexos costumam oferecer melhor desempenho em tarefas de visão computacional.

Por outro lado, apesar de atingirem melhor desempenho, os resultados mostram que as redes AlexNet e ResNet-34 estão associadas a valores consideravelmente maiores de pegada de carbono e consumo de energia do que as redes LeNet-5 e MobileNetV1. Conforme visto, existe uma diferença estatisticamente significativa entre os valores gerados por esses dois grupos para tais métricas. Ao mesmo tempo em que são mais econômicas do ponto vista energético e exigem menos recursos computacionais, as redes LeNet-5 e MobileNetV1 conseguem atingir um valor apenas ligeiramente inferior para as métricas de desempenho quando aplicadas à base MNIST. Assim, nesse caso, essas redes podem ser uma opção mais viável e sustentável do que as redes mais complexas, indo ao encontro do que é preconizado pela IA Verde.

Apesar de os experimentos computacionais terem focado no uso da base MNIST e de tais quatro arquiteturas de redes neurais para visão computacional, resultados similares podem ser atingidos considerando-se outras bases de dados, arquiteturas e tarefas, conforme evidenciado em trabalhos recentes da literatura [Strubell et al. 2019, Lenherr et al. 2021, Desislavov et al. 2023].

5. Conclusão

Este trabalho apresentou uma avaliação experimental entre diferentes arquiteturas de redes neurais convolucionais, considerando o desempenho, a eficiência energética e o custo computacional, para a base de dados MNIST. Os resultados mostraram que, apesar das redes AlexNet e ResNet-34 terem alcançado valores maiores em relação às métricas de avaliação de desempenho, tais redes são mais dispendiosas sob o ponto de vista energético. Por outro lado, as redes LeNet-5 e MobileNetV1 atingiram um desempenho ligeiramente inferior, mas estão associadas a valores de métricas de eficiência energética significativamente menores, conforme demonstrado pelo resultado dos testes de hipótese conduzidos.

A discussão sobre IA verde gira em torno da necessidade de desenvolver modelos que conciliem alto desempenho com sustentabilidade ambiental. Um caminho natural de trabalho futuro envolve a experimentação com outras bases de dados e arquiteturas de redes neurais, a fim de se expandirem os resultados obtidos e se reforçar o discurso preconizado pela IA Verde.

Referências

Aggarwal, C. C. (2023). *Neural Networks and Deep Learning: A Textbook*. Springer Publishing Company, Incorporated, 2nd edition.

Desislavov, R., Martínez-Plumed, F., and Hernández-Orallo, J. (2023). Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, 38:100857.

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nat. Mach. Intell.*, 2(8):423–425.

Douwes, C., Esling, P., and Briot, J.-P. (2021). Energy consumption of deep generative audio models. *arXiv preprint arXiv:2107.02621*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv.org*.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Hartwig, A. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv.org*.

Kelleher, J. D. (2019). *Deep Learning*. The MIT Press Essential Knowledge series, 1st edition.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Krohn, J. and Beyleveld, G. ang Bassens, A. (2020). *Deep Learning Illustrated. A Visual, Interactive Guide to Artificial Intelligence*. Addison Wesley Data & Analytics Series.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv.org*.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lenherr, N., Pawlitzek, R., and Michel, B. (2021). New universal sustainability metrics to assess edge intelligence. *Sustainable Computing: Informatics and Systems*, 31:100580.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green ai. *Communications of the ACM*, 63(12):54–63.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Wolff Anthony, L. F., Kanding, B., and Selvan, R. (2020). Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv.org*.