Inteligência Artificial Verde e Quantização em Modelos de Aprendizagem Profunda

Eduardo F. Gonçalves[1], Pedro Nuno de S. Moura[1], Daniel da S. Costa[1], Adriana C. F. Alvim[1]

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

{eduardo.f.goncalves, daniel.scosta}@edu.unirio.br,

{pedro.moura, adriana}@uniriotec.br

Abstract. *Concerns about the environment are playing an increasingly important role in society. Artificial Intelligence (AI), through Deep Learning (DL), offers useful tools for dealing with environmental issues. Despite the growing potential of DL models, they often require a lot of computing power, resulting in high levels of energy consumption and carbon footprint. Green AI aims to develop DL models that balance energy efficiency and performance in order to reduce environmental impact. This work focuses on quantization as a way of achieving the goals of Green AI, comparing how mixed-precision quantization impacts the energy consumption and performance of some neural networks for computer vision. The experimental results show that quantization can lead to the construction of more energy-efficient yet effective models, promoting a more sustainable approach.*

Resumo. *As preocupações com o meio ambiente têm assumido um papel cada vez maior na sociedade. A Inteligência Artificial (IA), por meio da Aprendizagem Profunda (AP), oferece ferramentas úteis no tratamento das questões ambientais. Apesar do crescente potencial dos modelos de AP, esses costumam demandar grande poder computacional, resultando em elevados níveis de consumo energético e pegada de carbono. A IA Verde se propõe a desenvolver modelos de AP que equilibrem eficiência energética e desempenho, a fim de diminuir o impacto ambiental. Este trabalho foca na quantização como uma forma de atingir os objetivos da IA Verde, comparando como a quantização de precisão mista impacta o consumo energético e o desempenho de algumas redes neurais para visão computacional. Os resultados experimentais apontam que a quantização pode levar à construção de modelos energeticamente mais eficientes e ainda assim efetivos, promovendo uma abordagem mais sustentável.*

1. Introdução

A sustentabilidade ambiental e a preocupação com as mudanças climáticas têm ganhado relevância no debate mundial. Em 2015, o Brasil e mais 192 Estados-membros da Organização das Nações Unidas assumiram o compromisso de promover o desenvolvimento equilibrado entre as dimensões econômica, social e ambiental por meio da aprovação da Agenda 2030, que inclui 17 Objetivos de Desenvolvimento Sustentável (ODS), com metas específicas para a eficiência energética e uso responsável dos recursos naturais (ODS 12) e ação contra a mudança global do clima (ODS 13) [Organização das Nações Unidas 2015].

A Inteligência Artificial (IA), sobretudo por meio dos modelos de Aprendizagem Profunda (AP), vem se destacando como uma ferramenta com potencial de auxiliar em algumas questões ambientais, a saber: combate às mudanças climáticas, gestão sustentável de recursos naturais e estratégias para uso eficiente de energia [Breder et al. 2024].

A pesquisa em IA tem levado a modelos de AP cada vez mais complexos e precisos, porém muitas vezes às custas de uma exigência significativa e crescente de recursos computacionais [Amodei and Hernandez 2018], sobretudo para realizar o treinamento de tais modelos. De fato, desde 2012 tem havido um crescimento exponencial no custo de treinamento de modelos de AP [Lenherr et al. 2021], resultando em uma demanda energética expressiva com consequente aumento da pegada de carbono. Tem-se, assim, um verdadeiro paradoxo entre a IA e a sustentabilidade, visto que, enquanto a IA pode ser uma ferramenta importante no desenvolvimento sustentável, ela também pode, ao mesmo tempo, impactar negativamente o meio ambiente [Breder et al. 2024]. Isso evidencia a necessidade de novas abordagens de IA que equilibrem eficiência energética e desempenho dos modelos, de forma a minimizar os impactos ambientais.

Foi nesse contexto que os conceitos de IA Vermelha (*Red AI*) e IA Verde (*Green AI*) foram propostos. A IA Vermelha refere-se a práticas que priorizam a maximização do desempenho dos modelos ao custo de impactos ambientais, ecônomicos e sociais, à medida que a IA Verde busca soluções que otimizem tanto o desempenho dos modelos quanto o consumo de recursos, fomentando práticas que reduzam o consumo de energia e a pegada de carbono [Schwartz et al. 2020].

Uma das abordagens mais promissoras para alcançar os objetivos da IA Verde é a quantização, que envolve a representação numérica de valores reais (contínuos) em valores discretos a fim de reduzir o número de *bits* necessários [Gholami et al. 2021]. No contexto de AP, essa abordagem corresponde a uma técnica que reduz a precisão numérica dos parâmetros de redes neurais, como, por exemplo, substituindo valores de ponto flutuante de 32 *bits* por representações de menor precisão, como valores de 16 ou 8 *bits*. Dessa forma, essa abordagem busca diminuir o custo computacional dos modelos, acelerar o tempo de processamento e reduzir significativamente o consumo energético.

Este trabalho tem como objetivo investigar a aplicação da quantização em modelos de AP, explorando o equilíbrio entre o uso de recursos computacionais, a eficiência energética e o desempenho dos modelos. Para isso, buscou-se avaliar experimentalmente como a quantização de precisão mista afeta o consumo energético e o desempenho de algumas arquiteturas de redes neurais para visão computacional, com foco na fase de inferência. Os resultados obtidos indicam que a quantização tem potencial de ser utilizada na obtenção de modelos mais eficientes e com menor impacto ambiental, ao mesmo tempo em que apresenta uma perda de desempenho mínima, consistindo assim em uma prática sustentável em IA.

O restante deste artigo está estruturado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados que abordam o custo computacional, a eficiência energética e a quantização de modelos de AP; já a Seção 3 descreve os materiais e métodos empregados nos experimentos computacionais; por sua vez, a Seção 4 detalha os resultados obtidos nos experimentos e os discute; e, por fim, a Seção 5 traz a conclusão e os trabalhos futuros.

2. Trabalhos Relacionados

As redes neurais voltaram a ter um renovado interesse dos pesquisadores a partir dos resultados surpreendentes obtidos pela rede AlexNet na competição sobre a base de dados ImageNet em 2012 [Krizhevsky 2009]. Amodei e Hernandez [Amodei and Hernandez 2018] observaram que, a partir desse marco, a quantidade de computação empregada no treinamento de redes neurais passou a dobrar a cada três ou quatro meses, dado o aumento progressivo no tamanho dos modelos.

Diante disso, [Schwartz et al. 2020] discutiram a preocupação com os altos custos da pesquisa em IA e propuseram alternativas mais sustentáveis que não implicassem no uso excessivo de recursos computacionais. Os autores apontaram que a IA Vermelha é menos acessível a pesquisadores com recursos limitados, além de ter um impacto ambiental significativo. Por outro lado, a IA Verde é mais inclusiva e prioriza a eficiência computacional como critério fundamental para a avaliação de modelos.

Considerando que os pesquisadores podem desempenhar um papel importante na redução do impacto ambiental por meio de projetos e implementações mais eficientes, os trabalhos de [Lacoste et al. 2019, Lottick et al. 2019, Wolff Anthony et al. 2020] se dedicaram a desenvolver ferramentas de quantificação da eficiência energética em implementações de modelos de Aprendizagem de Máquina. As ferramentas CodeCarbon [Lottick et al. 2019] e CarbonTracker [Wolff Anthony et al. 2020] utilizam interfaces de monitoramento para medir o consumo em CPUs e GPUs, enquanto a ferramenta MLCO2 [Lacoste et al. 2019] fornece uma estimativa das emissões de carbono a partir da informação da infraestrutura adotada.

O trabalho de [Ferraro et al. 2024] realizou uma avaliação experimental com modelos de AP de diferentes tamanhos para uma tarefa de visão computacional, considerando métricas de avaliação de desempenho e de eficiência energética. Para tal, adotaram a ferramenta CarbonTracker na aferição da pegada de carbono. Os autores demonstraram a viabilidade de se obterem modelos de AP que tenham menor custo computacional e que sejam mais econômicos energeticamente, ao passo que não apresentam uma diferença significativa de desempenho em relação a modelos mais dispendiosos.

As pesquisas conduzidas por [Desislavov et al. 2023] e por [Luccioni et al. 2024] concentraram-se na mensuração do consumo de energia durante a fase de inferência de diversos modelos destinados a tarefas de visão computacional e processamento de linguagem natural (PLN). Esses trabalhos argumentam que o custo de inferência de um modelo pode exceder o seu custo de treinamento, visto que, uma vez que esteja em produção, há um fator multiplicativo associado aos seus sucessivos usos para inferência.

Nesse sentido, o trabalho de [Luccioni et al. 2024] realizou uma comparação detalhada entre os custos de treino e de inferência em quatro modelos de linguagem e encontrou valores que indicam um custo de treino entre 200 a 590 milhões de vezes maior, em termos de consumo energético, do que o da inferência. Contudo, ao considerarem o dado de que o ChatGPT registrou mais de 1,7 bilhão de visitas em outubro de 2023, os autores concluíram que, devido ao volume intensivo de uso, os custos de inferência poderiam, potencialmente, ultrapassar os custos de treinamento em um período relativamente curto, que poderia ser de meses ou até mesmo de semanas.

Gholami *et al*. [Gholami et al. 2021] realizaram uma revisão da literatura sobre a

aplicação da quantização a modelos de AP, apresentando conceitos fundamentais, diferenciando métodos, abordagens e tipos de ajuste. Os autores destacaram o crescente consumo de recursos computacionais pelos modelos de AP e apontaram a quantização como uma possibilidade de tornar tais modelos mais eficientes, reduzindo a precisão numérica dos seus parâmetros sem comprometer significativamente o desempenho.

Rajput e Sharma [Rajput and Sharma 2024] apontaram que os diversos métodos de quantização podem impactar de formas diferentes a eficiência energética. Eles então propuseram um *benchmarking* comparativo de consumo de energia entre alguns métodos de quantização para um mesmo modelo de AP, que correspondeu ao Llama-2-7B. Os autores comprovaram que, sob o ponto de vista energético, existem métodos de quantização mais eficientes do que outros, podendo-se chegar a 200% de economia de energia comparativamente.

O trabalho de [Micikevicius et al. 2018] introduziu o conceito de quantização de precisão mista, que consiste na combinação, em um modelo de AP, de precisões numéricas de 32 e 16 *bits*. Os autores empregaram três métodos distintos de precisão mista em várias aplicações de AP, obtendo como resultado uma redução significativa do uso de memória e do tempo de treinamento, sem comprometer o desempenho dos modelos.

Finalmente, cabe também destaque ao trabalho de [Bouza et al. 2023] que conduziu uma extensa revisão e um teste de diversas ferramentas para estimativa de gasto energético e níveis de emissões no treinamento de modelos de AP, apontando o CodeCarbon como a melhor solução de software.

3. Materiais e Métodos

Adotou-se, neste trabalho, uma abordagem experimental para avaliar e comparar o efeito da quantização em diferentes arquiteturas de redes neurais voltadas à visão computacional, com foco na fase de inferência, considerando métricas de desempenho e eficiência energética. A seguir, os materiais e métodos empregados são descritos em detalhes.

3.1. Métricas

As métricas selecionadas têm como objetivo avaliar tanto o desempenho técnico dos modelos de AP quanto a sua sustentabilidade. Assim sendo, em relação ao desempenho, foram selecionadas métricas comumente adotadas na literatura para a avaliação de modelos voltados a problemas de classificação, a saber: acurácia, precisão, revocação e medida-F [Aggarwal 2023]. As suas definições são as seguintes: a acurácia corresponde à taxa de acerto de um modelo; já a precisão mede o quanto um modelo prediz corretamente exemplos de uma dada classe; a revocação mede a capacidade de um modelo de identificar corretamente exemplos de uma dada classe; e, por fim, a medida-F consiste na média harmônica entre a precisão e a revocação.

As métricas de eficiência energética adotadas neste trabalho seguem os princípios da IA Verde, sendo essenciais para mensurar a sustentabilidade dos modelos de AP e compreender o seu impacto ambiental. As métricas consideradas incluem a quantidade de energia utilizada durante a execução das tarefas, referente ao agregado de energia elétrica consumida por CPU, GPU e memória RAM, assim como a pegada de carbono, que é expressa pela quantidade de dióxido de carbono (CO_2) liberado na atmosfera oriundo desse consumo, e o próprio tempo de processamento tomado.

Para a coleta de tais métricas, utilizou-se a ferramenta CodeCarbon[1], proposta por [Lottick et al. 2019], que permite monitorar separadamente o consumo de energia de CPU, GPU e RAM. A ferramenta também estima as emissões de carbono equivalentes com base na localização geográfica do ambiente computacional usado e mede o tempo de processamento de trechos específicos de código.

Adicionalmente, a fim de refletir o custo computacional, foram coletados, para cada modelo, a sua quantidade de parâmetros treináveis e o seu número de FLOPs, que representa o montante de operações de ponto flutuante empregado na realização de uma determinada tarefa (e.g., uma inferência).

3.2. Conjunto de Dados

O conjunto de dados adotado nesta pesquisa foi o CIFAR-10[2], proposto por [Krizhevsky 2009], que é composto por 60.000 imagens coloridas e de baixa resolução (32x32 *pixels*). As imagens estão distribuídas em 10 classes. Cada classe contém 6.000 imagens, em que 5.000 são destinadas para treino e 1.000 para teste. Nos experimentos realizados, foi feita uma subdivisão nas imagens do conjunto de treino na proporção de 80% (40.000) para treino e 20% (10.000) para validação, além de um pré-processamento que compreendeu uma normalização dos canais RGB das imagens.

3.3. Arquiteturas Adotadas

Neste trabalho, foram utilizadas as seguintes arquiteturas: LeNet-5, AlexNet, GoogLeNet, VGG e ResNet. Elas foram escolhidas por representarem marcos significativos no desenvolvimento de redes neurais convolucionais (CNNs), sendo amplamente adotadas na literatura e em *benchmarks*. Foram realizadas adaptações nas arquiteturas para adequá-las ao número de classes do conjunto de dados e, nas camadas de entrada, para as dimensões das imagens (32x32 *pixels*). A Tabela 1 apresenta um resumo da quantidade de parâmetros e número de FLOPs dessas cinco arquiteturas.

Table 1. Quantidade de parâmetros e número de FLOPs das arquiteturas adotadas.

Rede	Parâmetros	FLOPs
LeNet-5	307.006	4.180.008
AlexNet	35.855.178	200.515.584
GoogLeNet	534.554	137.448.608
VGG-11	29.904.202	235.577.344
ResNet-34	21.282.122	1.163.507.200

A primeira proposta de CNN, a LeNet-5, foi desenvolvida por [LeCun et al. 1998] com o objetivo de realizar o reconhecimento automático de dígitos manuscritos. Ela é composta por camadas alternadas de convolução, *pooling* e camadas totalmente conectadas, permitindo a extração de características das imagens em diferentes níveis de abstração.

[1]https://codecarbon.io/

[2]https://www.cs.toronto.edu/∼ kriz/cifar.html

A AlexNet, proposta por [Krizhevsky et al. 2012], obteve resultados pioneiros no desafio *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) de 2012 e consistiu em um marco importante na AP. Sua arquitetura é mais complexa do que a LeNet-5, sendo composta por oito camadas convolucionais e três totalmente conectadas. Diversas técnicas atuais estavam presentes nessa arquitetura, como, por exemplo, o uso da função de ativação ReLU e *dropout*. Originalmente, foi treinada de forma paralela usando duas GPUs [Aggarwal 2023].

Por sua vez, a GoogLeNet, proposta por [Szegedy et al. 2015], se destacou por sua eficiência computacional e alto desempenho em tarefas de classificação de imagens. Seu sucesso se deve à utilização de módulos Inception. Cada módulo Inception é uma camada que permite a extração de informações das imagens, em diferentes níveis de granularidade, ao aplicar, paralelamente, filtros de tamanhos diferentes.

Já a VGG foi proposta por [Simonyan and Zisserman 2015]. Sua principal contribuição foi o aumento da profundidade da rede, graças à utilização de filtros menores (3×3), permitindo à rede aprender representações das imagens em mais níveis abstratos. Os autores testaram diferentes configurações variando a quantidade de camadas entre 11 (VGG-11) e 19 (VGG-19). Este trabalho usou a VGG-11.

Finalmente, a ResNet foi proposta por [He et al. 2015] e introduziu o conceito de aprendizado residual, através do uso de *skip connections*, que permitem que os dados de entrada de uma camada sejam copiados e utilizados em camadas seguintes, de forma não hierárquica. Essa modificação trouxe eficiência no treinamento e mitigou o problema do desaparecimento do gradiente em redes muito profundas. Este trabalho usou a ResNet-34, que apresenta 34 camadas.

3.4. Ambiente Computacional

O ambiente computacional adotado nos experimentos foi o seguinte: processador 13th Gen Intel(R) Core(TM) i7-13650HX, 16 GB DDR5 de memória RAM, placa de vídeo NVIDIA GeForce RTX 3050 6GB GDDR6 e sistema operacional Linux Mint 22.1 com kernel na versão 6.8.0. A implementação dos modelos foi realizada na linguagem de programação Python na versão 3.12, usando-se a biblioteca PyTorch[3] na versão 2.2. Para as medições de consumo energético e emissões de carbono, foi empregada a biblioteca CodeCarbon na versão 2.8.3. Os códigos desenvolvidos e os resultados dos experimentos encontram-se disponíveis publicamente em um repositório no GitHub[4].

3.5. Avaliação Experimental

A avaliação envolveu a experimentação com cada uma das cinco arquiteturas de redes neurais considerando duas variantes de precisão numérica: a versão original, com precisão de ponto flutuante de 32 *bits*, e uma versão quantizada com precisão mista utilizando o pacote *Automatic Mixed Precision*[5], em que pesos, ativações e dados de entrada utilizam ponto flutuante de meia precisão (16 *bits*), enquanto normalizações e cálculo da função de perda mantêm a precisão padrão (*32 bits*).

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

[4]https://github.com/dudu1626/wcama

[5]https://pytorch.org/docs/stable/amp.html

Nesta etapa de avaliação, houve a adoção de duas fases: treinamento e inferência. Na primeira fase, cada combinação de arquitetura de rede e variante de precisão numérica foi treinada dez vezes, registrando-se a perda e a acurácia de cada modelo em relação aos conjuntos de treino e de validação. Para cada combinação, foi escolhido o modelo que apresentou a maior acurácia para o conjunto de validação, a fim de ser usado na segunda fase. Nessa segunda fase, para cada combinação, o modelo escolhido foi utilizado para inferência em relação ao conjunto de teste, por dez vezes consecutivas, tomando-se as métricas de desempenho e de eficiência energética associadas.

No treinamento, a taxa de aprendizado foi de $0{,}001$ e foi adotada uma estratégia de ajuste dinâmico dessa, com uma paciência de cinco épocas e um fator de redução de $0{,}1$, monitorando-se a perda para o conjunto de validação. O otimizador escolhido foi o Adam e a função de perda correspondeu à entropia cruzada. Cada modelo foi treinado por um total de 50 épocas, utilizando um tamanho de lote de 32. Para garantir a reprodutibilidade, todos os modelos foram inicializados com pesos aleatórios utilizando uma semente fixa de valor igual a 158763.

4. Resultados Obtidos

A Tabela 2 apresenta a média e o desvio padrão dos valores das métricas de eficiência energética para cada combinação testada de arquitetura e variante de precisão numérica, considerando-se dez execuções da inferência para todo o conjunto de teste. É possível observar que, quanto maior o tamanho da rede, maior a redução nos valores das métricas proporcionada pela quantização. Enquanto a LeNet-5, que é a menor rede, não apresentou melhoria no tempo de execução, a ResNet-34, que é a maior, teve o seu tempo de processamento reduzido em $34{,}55\%$. Em relação ao consumo de energia e pegada de carbono, todos os modelos se beneficiaram: uma redução de $6{,}16\%$ para a LeNet-5 e redução de $32{,}96\%$ para a ResNet-34.

Table 2. Média e desvio padrão do tempo de processamento, gasto energético, em watts por hora, e emissão de carbono, em gramas, para a fase de inferência.

Arquitetura	Prec. Num.	Tempo (s)		Carbono (g)		Energia (W/h)	
		Média	DP	Média	DP	Média	DP
LeNet-5	32 *bits*	0.8686	0.0082	0.0033	0.0011	0.0336	0.0108
	16 *bits*	0.8717	0.0084	0.0031	0.0006	0.0315	0.0059
AlexNet	32 *bits*	1.4504	0.0122	0.0072	0.0007	0.0737	0.0073
	16 *bits*	1.3755	0.0097	0.0066	0.0008	0.0676	0.0084
GoogLeNet	32 *bits*	1.6603	0.0123	0.0086	0.0008	0.0875	0.0082
	16 *bits*	1.1084	0.0080	0.0068	0.0012	0.0687	0.0119
VGG-11	32 *bits*	1.7847	0.0076	0.0091	0.0015	0.0928	0.0150
	16 *bits*	1.5387	0.0072	0.0079	0.0014	0.0804	0.0143
ResNet-34	32 *bits*	5.8673	0.0111	0.0299	0.0036	0.3042	0.0365
	16 *bits*	3.8399	0.0083	0.0201	0.0032	0.2039	0.0326

A Tabela 3 exibe a média e o desvio padrão das métricas de desempenho para cada combinação testada de rede e variante de precisão numérica, para a fase de inferência. Observa-se que a ResNet-34 obteve os melhores resultados para as variantes de 32 *bits* e de 16 *bits* de precisão numérica, dentre todas as redes testadas.

Table 3. Média e desvio padrão das métricas de acurácia, precisão, revocação e medida-F para a fase de inferência.

Arquitetura	Prec. Num.	Acurácia		Precisão		Revocação		Medida-F	
		Média	DP	Média	DP	Média	DP	Média	DP
LeNet	32 *bits*	0.7303	0.0019	0.7305	0.0018	0.7303	0.0019	0.7301	0.0018
	16 *bits*	0.7372	0.0017	0.7374	0.0017	0.7372	0.0017	0.7371	0.0017
AlexNet	32 *bits*	0.8286	0.0016	0.8294	0.0017	0.8286	0.0016	0.8288	0.0016
	16 *bits*	0.8303	0.0020	0.8305	0.0020	0.8303	0.0020	0.8304	0.0020
GoogLeNet	32 *bits*	0.8424	0.0018	0.8421	0.0019	0.8424	0.0018	0.8421	0.0018
	16 *bits*	0.8427	0.0013	0.8424	0.0013	0.8427	0.0013	0.8425	0.0013
VGG-11	32 *bits*	0.8604	0.0015	0.8607	0.0015	0.8604	0.0015	0.8604	0.0015
	16 *bits*	0.8617	0.0014	0.8619	0.0014	0.8617	0.0014	0.8617	0.0014
ResNet-34	32 *bits*	0.8707	0.0014	0.8704	0.0014	0.8707	0.0014	0.8704	0.0014
	16 *bits*	0.8634	0.0009	0.8639	0.0010	0.8634	0.0009	0.8635	0.0010

Com a finalidade de demonstrar que a quantização é capaz de trazer diferença estatisticamente significativa para as métricas de eficiência energética, sem afetar significativamente as métricas de desempenho, os modelos foram separados em dois grupos: Grupo 1, constituído pelas redes com precisão numérica de 32 *bits* e Grupo 2, constituído pelas redes com precisão numérica de 16 *bits*.

Inicialmente, foi aplicado o teste de normalidade Shapiro-Wilk para os valores de tais métricas, considerando o Grupo 1 e o Grupo 2, e, como se verificou que as distribuições não eram normais, aplicou-se o teste de Mann-Whitney para a comparação entre os valores de tais métricas entre o Grupo 1 e o Grupo 2. Para as métricas de desempenho, não foram encontradas diferenças estatisticamente significativas, para uma significância de 5%, visto que todos os p-valores ficaram acima de $0{,}7$. Por outro lado, para as métricas de eficiência energética, foi possível observar diferença estatisticamente significativa, para uma significância de 6%, com todos os p-valores abaixo de $0{,}06$.

4.1. Discussão dos Resultados

Foi possível observar que a redução da precisão numérica de 32 *bits* para a precisão mista de 16 *bits* resultou em melhorias significativas na eficiência energética para todas as arquiteturas avaliadas, para a fase de inferência, com destaque para a ResNet-34, que apresentou uma redução de $32{,}96\%$ no consumo de energia e em emissões de carbono, além de uma redução de $34{,}55\%$ no seu tempo de processamento. Esses dados indicam que a quantização é capaz de diminuir o consumo de recursos computacionais, considerando a fase de inferência, que é um fator crítico para aplicações em produção e para a redução da pegada de carbono em larga escala, como apontado pelos trabalhos de [Desislavov et al. 2023, Luccioni et al. 2024].

Observando-se as métricas de desempenho, pode-se concluir que os resultados mostram estabilidade entre as versões quantizadas e não quantizadas das redes neurais, dado que a variação no desempenho foi mínima e não estatisticamente significativa. Isso sugere que a quantização pode ser implementada com segurança em modelos para produção sem comprometer a eficácia na tarefa de classificação de imagens. Tal constatação é crucial para a adoção das práticas da IA Verde, pois aponta que é possível manter a qualidade da saída dos modelos de AP ao mesmo tempo em que se reduz o seu impacto

ambiental.

A comparação entre diferentes arquiteturas nos permite inferir que a sensibilidade à quantização varia conforme a complexidade e arquitetura da rede. Enquanto a LeNet-5 não obteve melhorias no tempo de execução e apresentou diminuição modesta no consumo de energia e emissão de carbono (6,16%), a ResNet-34 mostrou reduções substanciais no tempo de processamento (34,55%) e nas emissões de carbono (32,96%).

5. Conclusão

Este trabalho demonstrou que a quantização de modelos de AP, especificamente na redução de precisão numérica de 32 *bits* para 16 *bits*, pode ser uma estratégia eficaz para promover a eficiência energética sem comprometer significativamente o desempenho de redes neurais convolucionais para a tarefa de classificação de imagens do conjunto de dados CIFAR-10. Os resultados experimentais alcançados com as arquiteturas LeNet-5, AlexNet, GoogLeNet, VGG-11 e ResNet-34 evidenciaram reduções no tempo de processamento, no consumo de energia e nas emissões de carbono para a fase de inferência dos modelos. Os resultados reforçam a viabilidade de aplicação da quantização como uma prática que permite atingir os objetivos preconizados pela IA Verde. Para trabalhos futuros, uma oportunidade se encontra na exploração de quantização em níveis mais baixos de precisão, assim como a combinação com outras técnicas de otimização de redes neurais, como poda ou destilação, para avaliar sinergias que possam aprimorar o equilíbrio entre eficiência e desempenho.

References

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

Amodei, D. and Hernandez, D. (2018). AI and Compute. `https://openai.com/index/ai-and-compute/`. Acesso em: 27 set. 2024.

Bouza, L., Bugeau, A., and Lannelongue, L. (2023). How to estimate carbon footprint when training deep learning models? a guide and review. *Environmental Research Communications*, 5(11):115014.

Breder, G., Brum, D., Dirk, L., and Ferro, M. (2024). O paradoxo da ia para sustentabilidade e a sustentabilidade da ia. In *Anais do V Workshop sobre as Implicações da Computação na Sociedade*, pages 105–116, Porto Alegre, RS, Brasil. SBC.

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.

Ferraro, V., Gullo, G., Costa, D., and Moura, P. (2024). Aprendizagem profunda e inteligência artificial verde: Caminhos para um futuro mais sustentável. In *Anais do XV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, pages 159–168, Porto Alegre, RS, Brasil. SBC.

Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., and Keutzer, K. (2021). A survey of quantization methods for efficient neural network inference. *CoRR*, abs/2103.13630.

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

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto.

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

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.

Lottick, K., Susai, S., Friedler, S. A., and Wilson, J. P. (2019). Energy usage reports: Environmental awareness as part of algorithmic accountability. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*.

Luccioni, S., Jernite, Y., and Strubell, E. (2024). Power hungry processing: Watts driving the cost of ai deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 85–99, New York, NY, USA. Association for Computing Machinery.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. (2018). Mixed precision training.

Organização das Nações Unidas (2015). Transformando nosso mundo: a agenda 2030 para o desenvolvimento sustentável. `https://tinyurl.com/agenda20230`. Acesso em: 12 mar. 2025.

Rajput, S. and Sharma, T. (2024). Benchmarking emerging deep learning quantization methods for energy efficiency. In *2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C)*, pages 238–242.

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

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*. arXiv:1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1–9.

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