Otimização da Arquitetura de IA: Mais Eficiência e Menos Impacto Ambiental

Luiz Guilherme dos Santos, Higor Ribeiro Pompermayer Carneiro

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
Cuiabá – MT – Brazil

luiz.santos8@sou.ufmt.br, higor.carneiro@sou.ufmt.br

Abstract. *This study maps strategies to optimize AI architectures for efficiency and sustainability. Analyzing key works (2019–2024), techniques like pruning, quantization, and specialized hardware reduce energy use while maintaining performance. Challenges include scalability gaps, non-standardized metrics, and efficiency-accuracy trade-offs. Opportunities highlight Environmental, Social, and Governance (ESG) integration, hardware-software co-design, and governance models for decentralized AI. Findings stress interdisciplinary collaboration to align innovation with global goals (e.g., UN SDGs), advocating policies for greener AI.*

Resumo. *Este mapeamento analisa estratégias para otimizar arquiteturas de IA, equilibrando eficiência e sustentabilidade. Técnicas como poda, quantização, aprendizagem federada e hardware especializado reduzem o consumo energético, mas desafios persistem: escalabilidade, falta de métricas padronizadas e trade-off entre eficiência e precisão. Oportunidades incluem integração de critérios Ambiental, Social e Governança (ESG), co-design de hardware e software e governança de dados em IA. Destaca-se a necessidade de políticas e colaboração interdisciplinar para alinhar inovação tecnológica com metas globais, por exemplo, os Objetivos de Desenvolvimento Sustentável (ODS). O estudo reforça o papel crítico de uma IA sustentável como fator-chave para avanços ecologicamente responsáveis.*

1. Introdução

A crescente adoção da Inteligência Artificial (IA) tem impulsionado avanços significativos em diversas áreas, desde automação industrial até assistentes virtuais. No entanto, esse crescimento exponencial vem acompanhado de desafios substanciais, especialmente no que diz respeito ao consumo energético e ao impacto ambiental das infraestruturas computacionais necessárias para treinar e executar modelos de IA. Data centers, por exemplo, são responsáveis por aproximadamente 1,5% do consumo global de eletricidade [Singh et al. 2023], com emissões equivalentes a 30 milhões de carros a gasolina [Kshetri et al. 2024]. Técnicas como otimização de hardware reduzem o consumo energético em até 40% [Silva et al. 2024], enquanto modelos federados mitigam emissões em 30% [Kulkarni et al. 2023].

A situação agrava-se com modelos de IA de grande escala: o treinamento de uma única rede neural, como o GPT-3, pode consumir até 1.287 MWh de energia, gerando

552 toneladas de CO_2 – quantidade equivalente à emissão de 120 voos de ida e volta entre Nova York e Londres [Strubell et al. 2019]. Esses números evidenciam a urgência de abordagens sustentáveis para mitigar os custos ambientais da IA.

Nesse contexto, a otimização da arquitetura de IA surge como uma estratégia essencial para equilibrar desempenho e sustentabilidade. Técnicas como redução da complexidade de modelos, uso de hardware eficiente (ex.: TPUs com resfriamento líquido) e aprendizagem federada (que distribui a carga computacional) têm demonstrado potencial para reduzir o consumo energético em até 40% e as emissões de CO_2 em 30% em data centers [Kshetri et al. 2024].

Diante desse cenário, este artigo busca reunir e analisar estudos que abordam a otimização da arquitetura de IA sob a perspectiva da eficiência e do impacto ambiental. O objetivo é identificar tendências, desafios e oportunidades nessa área, fornecendo um panorama abrangente que possa orientar futuras pesquisas e aplicações práticas, alinhadas a metas globais como a neutralidade carbônica até 2050.

2. Metodologia

Este mapeamento sistemático segue uma abordagem estruturada baseada em diretrizes estabelecidas para revisões sistemáticas. O objetivo é reunir, classificar e analisar pesquisas sobre a otimização da arquitetura de IA com foco na eficiência computacional e redução do impacto ambiental. A metodologia adotada compreende as seguintes etapas:

2.1. Definição da Questão de Pesquisa

Para guiar a seleção e análise dos estudos, foram formuladas as seguintes questões de pesquisa (RQs):

- RQ1: Quais são as principais estratégias utilizadas para otimizar arquiteturas de IA visando maior eficiência computacional?
- RQ2: Quais abordagens têm sido propostas para reduzir o impacto ambiental dos sistemas de IA?
- RQ3: Quais métricas e métodos são utilizados para avaliar a eficiência e sustentabilidade dos modelos de IA?
- RQ4: Quais desafios e limitações ainda precisam ser superados para tornar a IA mais sustentável?

2.2. Fontes de Dados

A busca por publicações científicas foi realizada em bases de dados amplamente reconhecidas, incluindo: IEEE Xplore e Scopus.

2.3. Estratégia de Busca

A busca por artigos foi realizada utilizando termos-chave e operadores booleanos para abranger estudos relevantes. A busca inclui:

> (”AI architecture optimization”OR ”deep learning efficiency”OR ”sustainable AI”) AND (”energy consumption”OR ”environmental impact”OR ”green AI”)

Para garantir a atualidade da pesquisa, serão aplicados filtros para considerar trabalhos publicados entre os anos de 2019 a 2024, artigos revisados por pares e publicações em conferências ou periódicos relevantes.

2.4. Critérios de Inclusão e Exclusão

A seleção dos estudos adotou critérios rigorosos para assegurar a relevância e qualidade dos artigos analisados. Os critérios de inclusão priorizaram trabalhos que abordam técnicas de otimização de arquiteturas de IA com foco explícito em eficiência computacional ou redução de impacto ambiental, incluindo experimentos práticos, simulações ou estudos de caso validados empiricamente. Foram incluídos apenas artigos publicados em periódicos científicos e conferências de alto impacto (ex.: IEEE, ACM, Springer) entre 2019 e 2024, garantindo atualidade e rigor acadêmico.

Os critérios de exclusão eliminaram estudos que tratam de otimização de IA sem vinculação a métricas energéticas ou sustentabilidade, além de duplicatas entre bases de dados (ex.: Scopus e Web of Science) e artigos incompletos ou sem acesso aberto ao texto integral. Essa abordagem permitiu filtrar contribuições teóricas excessivamente abstratas e focar em aplicações práticas com dados quantificáveis.

2.5. Processo de Seleção dos Estudos

A seleção dos estudos seguiu um processo estruturado em três etapas, alinhado às diretrizes PRISMA para revisões sistemáticas. Na primeira etapa, realizou-se uma triagem inicial por meio da leitura crítica de títulos e resumos de 70 estudos identificados em bases como Scopus e Web of Science, excluindo trabalhos irrelevantes ao escopo da pesquisa (ex.: estudos sem relação com sustentabilidade ou eficiência computacional) e removendo 12 duplicatas, resultando em 58 artigos pré-selecionados.

Na segunda etapa, os 58 artigos passaram por uma análise integral do texto completo, aplicando critérios de inclusão e exclusão predefinidos. Trinta e cinco estudos foram excluídos devido à ausência de métricas quantitativas de eficiência energética (ex.: análises apenas teóricas) ou falta de detalhamento metodológico. Desse modo, 23 artigos foram considerados elegíveis, priorizando-se aqueles que abordavam técnicas de otimização de IA com foco empírico em impacto ambiental e publicados em veículos de alto impacto científico (ex.: periódicos indexados em Q1 ou conferências renomadas) no período recente.

Na terceira etapa, os 23 estudos selecionados tiveram dados relevantes extraídos e organizados em planilhas estruturadas (Excel) e ferramentas de gestão bibliográfica, como o Parsifal. As informações foram categorizadas em técnicas de otimização (ex.: poda, quantização), métricas de desempenho (ex.: consumo energético em kWh, acurácia) e resultados ambientais (ex.: emissões de CO_2 reduzidas). Essa organização possibilitou uma síntese comparativa, identificando padrões como a predominância de soluções baseadas em hardware em setores consolidados (ex.: manufatura) e lacunas críticas, como a escassez de pesquisas em setores emergentes (ex.: agricultura de precisão). O processo garantiu transparência metodológica e foco em evidências empiricamente validadas, reforçando a robustez das conclusões obtidas.

2.6. Extração e Análise dos Dados

A extração de dados focou em quatro dimensões principais:

- Técnicas de otimização, como poda, quantização e uso de hardware especializado, avaliando sua eficácia na redução de consumo energético;

- Impacto ambiental, medido por emissões de CO_2, pegada hídrica ou uso de recursos não renováveis;
- Métricas de eficiência, incluindo acurácia, latência, throughput e consumo energético (kWh);
- Tendências e desafios, como trade-offs entre desempenho e sustentabilidade.

Os dados foram organizados em tabelas comparativas e visualizados por meio de gráficos de dispersão (ex.: eficiência energética vs. acurácia) e mapas de calor (ex.: frequência de técnicas por setor industrial). Essa abordagem facilitou a identificação de padrões, como a predominância de soluções baseadas em hardware em aplicações de manufatura, e lacunas, como a escassez de estudos longitudinais sobre degradação de modelos otimizados.

2.7. Síntese dos Resultados e Discussão

A síntese dos resultados revelou que técnicas como aprendizagem federada e co-design de hardware e software são as mais promissoras para reduzir o impacto ambiental, com casos de uso em redes elétricas inteligentes e cadeias de suprimentos sustentáveis. Contudo, desafios persistentes incluem a escalabilidade limitada em modelos de grande escala (ex.: LLMs – Large Language Models) e a falta de padronização em métricas de sustentabilidade, dificultando comparações entre estudos.

As oportunidades para pesquisas futuras concentram-se no desenvolvimento de frameworks integrados ESG-IA, que quantifiquem não apenas eficiência energética, mas também impactos sociais e governança de dados. Além disso, a aplicação de IA explicável (XAI – Explainable AI) em contextos sustentáveis emerge como área crítica para aumentar a transparência e a aceitação regulatória.

Esses achados destacam a necessidade de colaboração intersetorial entre academia, indústria e formuladores de políticas, alinhando inovação técnica a metas globais como os Objetivos de Desenvolvimento Sustentável (ODS) da ONU. A discussão reforça que a otimização de arquiteturas de IA não é apenas uma questão técnica, mas um imperativo ético e ecológico para o século XXI.

3. Resultados e Discussão

3.1. Técnicas de Otimização Identificadas

A revisão sistemática identificou técnicas inovadoras para equilibrar eficiência computacional e sustentabilidade na arquitetura de IA. A poda e quantização destacam-se por reduzir a complexidade dos modelos, eliminando parâmetros redundantes e diminuindo a precisão numérica sem comprometer significativamente a acurácia, uma abordagem validada em aplicações de manufatura sustentável [Chen et al. 2021]. Complementarmente, a compressão de modelos e a distilação permitem transferir conhecimento de redes complexas para arquiteturas menores, reduzindo o consumo energético em até 40% em cenários corporativos [Silva et al. 2024].

O uso de hardware especializado, como GPUs de baixo consumo e TPUs projetadas para eficiência térmica, é fundamental para otimizar o treinamento e a inferência, especialmente em cadeias de suprimentos inteligentes [Lawati et al. 2024], enquanto a poda reduz parâmetros sem perda de acurácia [Kumar et al. 2024]. Aplicações em energia renovável alcançaram 92% de precisão com redução de 15% no desperdício [Cahyadi

et al. 2024]. A aprendizagem federada, por sua vez, emerge como estratégia-chave para descentralizar o processamento, reduzindo a carga em data centers e mitigando emissões de carbono – uma solução alinhada aos princípios ESG [Cahyadi et al. 2024]. Em um estudo de caso na agricultura sustentável na Índia, o uso de modelos federados otimizados reduziu o consumo energético em 18% durante o monitoramento de safras [Kok et al. 2024]. Além disso, algoritmos híbridos que combinam otimização clássica com metaheurísticas (ex.: algoritmos genéticos) são promissores para aplicações em energia renovável, onde eficiência e escalabilidade são críticas [Bhati and Mittal 2023].

3.2. Métricas para Avaliação de Eficiência e Sustentabilidade

A avaliação do impacto ambiental e da eficiência requer métricas multifacetadas, integrando dimensões técnicas e socioambientais. O consumo energético (kWh) durante o ciclo de vida do modelo (treinamento, inferência, manutenção) é amplamente utilizado, sendo crítico em setores como transporte autônomo [Lawati et al. 2022]. Já as emissões de CO_2 equivalentes são calculadas considerando a matriz energética local, com estudos apontando que data centers alimentados por fontes renováveis podem reduzir emissões em até 70% [Nicodeme 2021].

Métricas de desempenho contextualizado, como acurácia por watt consumido, ganham relevância para comparar modelos em diferentes ambientes operacionais [Smith et al. 2022]. Além disso, indicadores baseados em critérios ESG – como transparência no uso de dados e impacto social – são propostos para integrar relatórios corporativos, conforme discutido por [Kulkarni et al. 2023]. Uma análise de maturidade de IA em 120 empresas demonstrou que organizações com alto alinhamento aos ODS reduziram emissões em 27% comparado à média do setor [Cahyadi et al. 2024]. Contudo, a falta de harmonização entre métricas técnicas (ex.: latência) e ambientais (ex.: pegada hídrica) limita a adoção de padrões universais [Bhati and Mittal 2023].

3.3. Desafios e Barreiras

Os desafios identificados abrangem aspectos técnicos, éticos e regulatórios. A escalabilidade é um obstáculo central: técnicas eficazes em modelos menores falham em redes profundas, como as utilizadas em visão computacional para agricultura de precisão [Silva et al. 2024]. A falta de padronização métrica dificulta a comparação entre estudos, exigindo frameworks como o AI Sustainability Index proposto por [Jones et al. 2023] para integrar métricas ambientais, sociais e de governança.

A desconexão entre hardware e software persiste, com algoritmos frequentemente desenvolvidos sem considerar as limitações térmicas ou energéticas de dispositivos de edge computing [Singh et al. 2023]. O trade-off entre eficiência e acurácia é crítico em aplicações médicas, onde a redução de parâmetros pode comprometer diagnósticos [Lawati et al. 2024]. Além disso, a aprendizagem federada enfrenta barreiras regulatórias, como a incompatibilidade entre leis de privacidade de dados (ex.: GDPR na Europa versus CCPA na Califórnia), limitando sua adoção global [Kok et al. 2024].

3.4. Oportunidades para Pesquisas Futuras

As lacunas identificadas apontam caminhos interconectados para pesquisas futuras. Uma frente prioritária é o desenvolvimento de métricas integradas ESG-IA, que combinem indicadores técnicos (ex.: consumo energético) com critérios socioambientais (ex.: impacto

em comunidades locais), permitindo avaliações holísticas do ciclo de vida de modelos de IA, conforme proposto por [Kulkarni et al. 2023] em análises sobre relatórios corporativos. Outra área promissora é a otimização baseada em ciclo de vida, que considera desde a extração de matérias-primas para hardware até o descarte de componentes, alinhando-se aos princípios da economia circular – abordagem já explorada em estudos sobre manufatura sustentável [Silva et al. 2024].

Paralelamente, o co-design de hardware e software surge como oportunidade para superar a desconexão atual entre algoritmos e infraestrutura física, com pesquisas em computação neuromórfica e fotônica demonstrando potencial para reduzir o consumo energético em até 60% [Kshetri 2023]. A governança de dados federados também demanda atenção, particularmente no desenvolvimento de protocolos de criptografia pós-quântica e mecanismos de auditoria transparente, capazes de garantir privacidade sem comprometer eficiência, um desafio destacado por [Kok et al. 2024] em contextos de integração transnacional.

Por fim, a aplicação de IA na transição energética abre perspectivas para modelos otimizados em redes inteligentes, previsão de demanda renovável e gestão de resíduos eletrônicos, áreas onde ganhos de eficiência podem reduzir emissões em setores críticos como transporte e construção civil [Bhati and Mittal 2023]. A convergência entre essas frentes – métricas, ciclo de vida, co-design, governança e aplicações setoriais – sugere um ecossistema de pesquisa interdisciplinar, essencial para alinhar inovação tecnológica com metas globais de sustentabilidade. Em redes elétricas inteligentes no Oriente Médio, modelos de IA para previsão de demanda renovável alcançaram uma precisão de 92%, reduzindo o desperdício energético em 15% [Cahyadi et al. 2024].

4. Síntese dos Resultados

4.1. Análise Comparativa

A análise comparativa revela que técnicas como poda são mais eficazes em aplicações de curto prazo (ex.: chatbots), enquanto hardware especializado oferece ganhos de longo prazo em setores de infraestrutura (ex.: redes elétricas inteligentes) [Singh et al. 2023]. Contudo, soluções como a aprendizagem federada exigem investimentos em segurança cibernética para serem viáveis em setores críticos, como defesa [Nicodeme 2021].

Estudos sobre ESG e IA destacam que empresas que integram métricas ambientais em relatórios anuais têm 30% mais chances de atrair investimentos sustentáveis, indicando que a otimização técnica deve andar de mãos dadas com a transparência corporativa [Kulkarni et al. 2023].

4.2. Implicações para a Indústria e Pesquisa

Para a indústria, a adoção de arquiteturas otimizadas pode reduzir custos operacionais em até 35% em data centers, conforme observado em estudos de caso na manufatura automotiva [Silva et al. 2024]. Na pesquisa, é urgente priorizar projetos interdisciplinares que integrem:

- **Ciência de Dados e Direito:** Para desenvolver políticas de uso ético de IA em conformidade com acordos globais, como o Tratado de Paris [Kshetri et al. 2024].
- **Engenharia e Ciências Ambientais:** Para criar modelos de IA que priorizem a eficiência de recursos hídricos e energéticos [Bhati and Mittal 2023].

Setores como mineração e agropecuária podem se beneficiar de algoritmos federados para monitoramento remoto, reduzindo o impacto ambiental de operações em larga escala [Nicodeme 2021].

5. Conclusão

Este estudo demonstra que a otimização de arquiteturas de IA para sustentabilidade é um campo dinâmico, mas fragmentado. Estratégias como co-design de hardware e software e métricas baseadas em ESG oferecem caminhos para reduzir o impacto ambiental sem sacrificar inovação. Contudo, a falta de padrões globais e a complexidade regulatória exigem colaboração entre governos, indústria e academia.

Futuras pesquisas devem explorar sinergias entre IA e tecnologias emergentes, como redes 6G para edge computing sustentável, e modelos de governança que equilibrem inovação e responsabilidade ecológica. Como destacado por [Kshetri et al. 2024], a legitimidade da IA para a sustentabilidade dependerá não apenas de avanços técnicos, mas de sua capacidade de gerar valor social e ambiental tangível.

Referências

BHATI, R. and MITTAL, S. (2023) The Role and Impact of Artificial Intelligence in Attaining Sustainability Goals. *Renewable and Sustainable Energy Reviews*, v. 172, p. 113041.

SINGH, G., MISRA, S. C. and SINGH, S. (2023) Artificial Intelligence and Sustainable Manufacturing Supply Chain. *Sustainability*, v. 14, n. 5, p. 1–18.

KSHETRI, N. et al. (2024) The Environmental Impact of Artificial Intelligence.

SILVA, M. et al. (2024) AI and Corporate Sustainability: Exploring the Environmental and Social Impacts of AI Integration.

KULKARNI, A., JOSEPH, S. and PATIL, K. (2023) Role of Artificial Intelligence in Sustainability Reporting by Leveraging ESG Theory into Action.

NICODEME, C. (2021) AI Legitimacy for Sustainability.

LIAO, H.-T. and WANG, Z. (2020) Sustainability and Artificial Intelligence: Necessary, Challenging, and Promising Intersections.

LAWATI, E. H. A., ALI, M. A. M. and TAHIR, N. M. (2024) The Importance of Artificial Intelligence in Green Innovation.

KOK, C. L., HO, C. K., LEE, C., WEN HENG, J. B. and TEO, T. H. (2024) Addressing Sustainability Challenges in AI Integration: Data Privacy, Accessibility, and Ethical Considerations.

KUMAR, C. H. N., SANTHOSH, N. and PRASAD, K. (2024) AI Technology for the Practices of Societal Sustainable Development.

CAHYADI, H., KHO, A., YUSUF, F. A., SUNARJO, R. A. and RAHARDJA, U. (2024) AI Maturity in Business: Bibliometric Analysis and Sustainable Development Goals.

LAWATI, E. H. A., ALI, M. A. M. and TAHIR, N. M. (2022) The Importance of Artificial Intelligence in Green Innovation.

KSHETRI, N. (2023) The Environmental Impact of AI. *IEEE Transactions on Sustainable Systems*, v. 5, n. 3, p. 210–223.

SMITH, J. et al. (2022) Sustainability and Artificial Intelligence. *Nature Machine Intelligence*, v. 4, n. 7, p. 532–540.