# O USO DE INTELIGÊNCIA ARTIFICIAL NA OTIMIZAÇÃO DE ARQUITETURAS DE HARDWARE: melhorando o processamento, reduzindo custos e ampliando o acesso a tecnologias de processamento neural

## *USING ARTIFICIAL INTELLIGENCE TO OPTIMIZE HARDWARE ARCHITECTURES: improving processing, reducing costs, and expanding access to neural processing technologies*

Vitor Bruno Teodoro[1]; Alberane Lúcio Thiago da Cunha[2]

## RESUMO

Este trabalho investiga o uso de técnicas de inteligência artificial (IA) para a otimização de arquiteturas de *hardware*, com foco em processadores com núcleos neurais dedicados. O objetivo central é compreender como métodos de IA, como aprendizado de máquina, algoritmos genéticos e aprendizado por reforço, podem contribuir para aprimorar o desempenho computacional, a eficiência energética e reduzir os custos de fabricação desses processadores. A pesquisa adota uma abordagem qualitativa, exploratória e descritiva, fundamentada em revisão bibliográfica sistematizada e análise comparativa de *benchmarks* de desempenho e eficiência de diferentes plataformas de *hardware*. Também são considerados estudos de caso que demonstram a aplicação prática dessas técnicas no *design* e manufatura de processadores neurais. Os resultados apontam que o uso de IA no processo de desenvolvimento de *hardware* permite acelerar ciclos de projeto, otimizar o uso de recursos e reduzir falhas, contribuindo para o desenvolvimento de soluções mais acessíveis e sustentáveis. Além disso, a integração de IA no *design* de *hardware* favorece a democratização do acesso a tecnologias de alto desempenho, impactando positivamente setores como saúde, indústria, segurança, educação e economia digital. Conclui-se que a aplicação dessas técnicas representa um caminho promissor para superar os desafios de

[1] Aluno do curso de Sistemas de Informação do Centro Universitário do Sul de Minas. Email: brunoteodoro23@hotmail.com

[2] Coordenador do curso de Sistemas de Informação do Centro Universitário do Sul de Minas e Orientador do TCC. Email: alberane.cunha@professor.unis.edu.br

custo e complexidade dos processadores neurais, fortalecendo sua presença em dispositivos móveis, sistemas embarcados e ambientes corporativos.

**Palavras-chave:** Inteligência Artificial. Arquitetura de *Hardware*. Processadores Neurais. Otimização de Desempenho. Aprendizado de Máquina.

## *ABSTRACT*

*This study investigates the use of artificial intelligence (AI) techniques to optimize hardware architectures, focusing on processors with dedicated neural cores. The main objective is to understand how AI methods such as machine learning, genetic algorithms, and reinforcement learning can contribute to improving computational performance, energy efficiency, and reducing manufacturing costs of these processors. The research adopts a qualitative, exploratory, and descriptive approach, based on a systematic literature review and a comparative analysis of performance and efficiency benchmarks from different hardware platforms. Case studies demonstrating the practical application of these techniques in the design and manufacturing of neural processors are also considered. The results indicate that the use of AI in hardware development accelerates design cycles, optimizes resource utilization, and reduces failures, contributing to the development of more accessible and sustainable solutions. Furthermore, integrating AI into hardware design promotes the democratization of access to high-performance technologies, positively impacting sectors such as healthcare, industry, security, education, and the digital economy. It is concluded that the application of these techniques represents a promising path to overcoming the cost and complexity challenges of neural processors, strengthening their presence in mobile devices, embedded systems, and corporate environments.*

***Keywords:*** *Artificial Intelligence. Hardware Architecture. Neural Processors. Performance Optimization. Machine Learning.*

## 1 INTRODUÇÃO

Este trabalho investiga o uso de técnicas de Inteligência Artificial para a otimização de arquiteturas de *hardware*, com foco na melhoria de desempenho, eficiência energética e redução de custos de fabricação. A pesquisa aborda algoritmos de aprendizado de máquina aplicados ao *design* de circuitos, explorando abordagens como redes neurais, otimização baseada em algoritmos genéticos e aprendizado por reforço. O estudo será limitado a aplicações em *hardware* especializado, analisando casos reais e simulações para demonstrar os impactos das técnicas propostas.

O desenvolvimento e a consolidação de novas tecnologias baseadas em inteligência artificial têm promovido avanços significativos no campo da arquitetura de *hardware*, especialmente no que se refere ao aprimoramento de processadores com núcleos neurais integrados. Esses núcleos especializados são projetados para executar, de forma otimizada, operações relacionadas ao treinamento e à inferência de modelos de aprendizado de máquina, conferindo maior eficiência energética e redução de latência em comparação com processadores tradicionais ou unidades de processamento gráfico (GPUs). Nesse contexto, o presente trabalho tem como objetivo investigar a integração entre inteligência artificial e arquiteturas de *hardware*, com ênfase no estudo dos processadores que incorporam núcleos neurais dedicados. A pesquisa buscará compreender como tais processadores podem contribuir para a melhoria do desempenho computacional, a redução de custos operacionais e a ampliação do acesso a tecnologias de inteligência artificial de alto desempenho.

Para delimitar o escopo da pesquisa, este estudo é conduzido com foco em processadores contemporâneos que incorporam núcleos neurais dedicados, amplamente empregados em dispositivos móveis, sistemas embarcados e servidores voltados à execução de tarefas de inteligência artificial. A metodologia adotada contemplará uma revisão bibliográfica sistematizada, visando mapear a evolução histórica e tecnológica dessas arquiteturas, bem como uma análise comparativa entre diferentes plataformas e fabricantes, com base em *benchmarks* de desempenho em cenários de inferência e treinamento de modelos de IA. Ademais, o estudo busca identificar de que modo a utilização de técnicas de inteligência artificial pode contribuir para a otimização do próprio processo de *design* e manufatura desses processadores, com vistas à melhoria da eficiência térmica e energética. O recorte temporal da pesquisa abrange o período a partir de 2020, marco em que os processadores com núcleos neurais especializados passaram a ser amplamente utilizados em diferentes segmentos da indústria, enquanto o recorte espacial

concentrar-se-á em tecnologias aplicadas tanto em dispositivos de consumo quanto em soluções corporativas.

O crescimento do uso de inteligência artificial tem impulsionado a necessidade de processadores especializados, capazes de oferecer maior eficiência computacional e energética. No entanto, a otimização dessas arquiteturas de *hardware* ainda enfrenta desafios relacionados à complexidade do *design*, ao alto custo de fabricação e à necessidade de adaptação para diferentes cargas de trabalho. Nesse contexto, a questão central que este trabalho busca responder é: como as técnicas de inteligência artificial podem contribuir para a otimização do *design* e da eficiência de processadores neurais, promovendo melhor desempenho, redução de custos e ampliação do acesso à tecnologia?

Acredita-se que a aplicação de técnicas de inteligência artificial, como aprendizado de máquina, algoritmos genéticos e aprendizado por reforço, pode contribuir significativamente para a otimização de arquiteturas de *hardware*, especialmente processadores com núcleos neurais, promovendo ganhos em desempenho, eficiência energética e redução de custos, tornando tais tecnologias mais acessíveis e amplamente utilizadas.

## 2 REFERENCIAL TEORICO

### 2.1 Inteligência Artificial: Conceitos e Aplicações

A Inteligência Artificial (IA) refere-se à capacidade de sistemas e máquinas simularem processos cognitivos humanos, como percepção, aprendizado, raciocínio e tomada de decisão. Segundo Russell e Norvig (2021), IA é o estudo de agentes inteligentes, sendo definidos como qualquer dispositivo que percebe seu ambiente e executa ações que maximizam suas chances de sucesso em algum objetivo.

Dentre os principais campos da IA destacam-se o aprendizado de máquina (*machine learning*), o aprendizado profundo (*deep learning*) e os algoritmos evolutivos, que, além de promoverem avanços na automação de processos, também têm sido aplicados na otimização de projetos de hardware. Para Goodfellow, Bengio e Courville (2016), o aprendizado de máquina permite que sistemas aprendam automaticamente padrões complexos sem serem explicitamente programados, tornando-se essencial para o avanço tecnológico atual.

Empresas líderes, como NVIDIA e Google, já aplicam esses conceitos em seus produtos, como na série NVIDIA GeForce RTX 50 (2025), que integra núcleos de IA especializados para acelerar processos de inferência. Essa integração entre IA e *hardware* inaugura uma nova era de eficiência computacional, na qual o próprio sistema é capaz de se adaptar e otimizar suas operações. Assim, a utilização de IA na otimização de *hardware* representa um avanço estratégico para a redução de custos, o aumento do desempenho e a ampliação do acesso a tecnologias de alto processamento.

## 2.2 Arquitetura de *Hardware* Otimizada para IA

A crescente demanda por IA impôs a necessidade de desenvolvimento de arquiteturas de *hardware* específicas, que sejam capazes de oferecer alto desempenho com menor consumo energético. Sze *et al*. (2017) destacam que arquiteturas tradicionais, como CPUs e GPUs, embora sejam versáteis, não são ideais para as cargas de trabalho exigidas pelos modelos de redes neurais profundas, principalmente devido ao elevado consumo energético e à limitação em paralelismo específico.

Diante disso, surgiram os Núcleos de Processamento Neural (NPUs), projetados para acelerar operações típicas de IA, como multiplicações e somas de matrizes, essenciais para o funcionamento de redes neurais. Esses processadores especializados oferecem, segundo Sze *et al*. (2017), "uma melhoria significativa na eficiência energética, na redução de latência e no custo operacional em comparação às arquiteturas tradicionais".

Empresas como Google, com seus TPUs, e Apple, com o *Neural Engine* presente nos seus processadores móveis, são exemplos claros da adoção de NPUs como solução de *hardware* voltada à IA.

## 2.3 A Inteligência Artificial no Processo de Otimização de *Hardware*

A utilização da própria IA para otimizar o *design* de *hardware* é uma tendência crescente. Mirhoseini (2021) demonstraram que algoritmos de aprendizado por reforço foram capazes de superar engenheiros humanos no problema de colocação de blocos de circuitos integrados, tarefa

crucial no projeto de chips. De acordo com os autores, “o método reduziu o tempo de desenvolvimento de semanas para horas, além de produzir *designs* mais eficientes”.

Essa abordagem, conhecida como *AI for Hardware Design*, permite que o ciclo de desenvolvimento de processadores seja acelerado, ao mesmo tempo em que melhora o desempenho, a eficiência energética e reduz custos.

## 2.4 Técnicas de IA Aplicadas à Otimização de *Hardware*

As técnicas de Inteligência Artificial aplicadas à otimização de *hardware* buscam aprimorar o desempenho, a eficiência energética e a redução de custos em sistemas computacionais. Por meio de algoritmos de aprendizado de máquina e redes neurais, é possível automatizar o *design*, o ajuste de parâmetros e o gerenciamento de recursos de *hardware*. Essa integração promove arquiteturas mais inteligentes, adaptáveis e eficientes para diferentes demandas tecnológicas.

### 2.4.1 Aprendizado de Máquina e Aprendizado Profundo

O aprendizado de máquina, conforme Goodfellow, Bengio e Courville (2016), consiste em métodos estatísticos que permitem que sistemas “aprendam” a realizar tarefas a partir de dados, sem serem explicitamente programados. No contexto de *hardware*, essas técnicas são aplicadas para prever falhas, otimizar *layouts* de circuitos, reduzir consumo energético e acelerar processos de desenvolvimento.

### 2.4.2 Aprendizado por Reforço

O aprendizado por reforço permite que agentes aprendam estratégias ideais por meio de interação com o ambiente, maximizando recompensas cumulativas. No contexto de *design* de *chips*, Mirhoseini. (2021) demonstrou que essa abordagem pode ser aplicada no posicionamento de blocos de circuitos, reduzindo o tempo de projeto e melhorando métricas como latência, área e consumo de energia.

### 2.4.3 Algoritmos Genéticos na Otimização

Diferente do aprendizado por reforço ou profundo, os Algoritmos Genéticos (GAs) são técnicas de busca e otimização inspiradas na teoria da evolução biológica e seleção natural. Eles operam sobre uma "população" de soluções candidatas para um determinado problema.

No contexto do design de hardware, os GAs são particularmente úteis para explorar espaços de busca muito amplos e complexos. Por exemplo, no design de um layout de circuito integrado (similar ao problema de posicionamento de blocos), um GA pode testar milhares de configurações (indivíduos), avaliando cada uma através de uma "função de aptidão" (que pode medir desempenho, área do chip ou consumo energético).

As melhores soluções são "selecionadas" para "reprodução" (combinando características através do crossover) e "mutação" (introduzindo pequenas alterações aleatórias), gerando uma nova população de soluções potencialmente melhores. Esse processo é repetido por várias gerações até que uma solução ótima ou satisfatória seja encontrada. Os GAs são muito eficientes em problemas de otimização multiobjetivo, onde é preciso balancear métricas concorrentes, como desempenho e custo (DEB et al., 2002).

## 2.5 Processadores Neurais: Características e Benefícios

Os processadores neurais são arquiteturas desenvolvidas para acelerar operações associadas a modelos de IA, especialmente redes neurais profundas. Segundo Sze et al. (2017), tais processadores são otimizados para:

a) Alta paralelização de operações;
b) Processamento de dados com menor precisão (como FP16, INT8), reduzindo consumo energético;
c) Eficiência térmica e energética superior a CPUs e GPUs em tarefas de inferência.
d) Essa classe de processadores está presente em diversos setores, como dispositivos móveis, sistemas embarcados, automóveis autônomos e soluções de *edge computing*, permitindo que aplicações de IA sejam executadas localmente com alta eficiência.

## 2.6 Desafios na Otimização de Arquiteturas de *Hardware* para IA

Embora os avanços sejam significativos, a otimização de *hardware* para IA enfrenta desafios consideráveis. Russell e Norvig (2021) destacam que, além da complexidade do *design*, os custos de fabricação de chips especializados ainda são elevados, exigindo soluções que consigam balancear desempenho, consumo energético e custo.

Outro desafio é a necessidade de projetar *hardware* flexível, capaz de atender a diferentes cargas de trabalho, dado que modelos de IA estão em constante evolução. Além disso, questões relacionadas à sustentabilidade ambiental se tornam cada vez mais relevantes, considerando o alto impacto energético dos centros de dados e da fabricação de *chips* (Sze *et al*., 2017).

## 2.7 Estado da Arte e Perspectivas Futuras

A partir de 2020, observa-se uma intensificação na adoção de processadores com núcleos neurais dedicados, refletindo a crescente demanda por eficiência energética e desempenho em aplicações de IA. As perspectivas futuras incluem o desenvolvimento de arquiteturas heterogêneas, combinando NPUs, GPUs e CPUs, além da automação quase completa dos processos de *design* de *hardware* através de IA (Mirhoseini., 2021).

Espera-se, também, a democratização do acesso a tecnologias de IA de alto desempenho, impulsionada pela redução dos custos de desenvolvimento e fabricação, beneficiando tanto o mercado de consumo quanto soluções corporativas.

# 3 METODOLOGIA

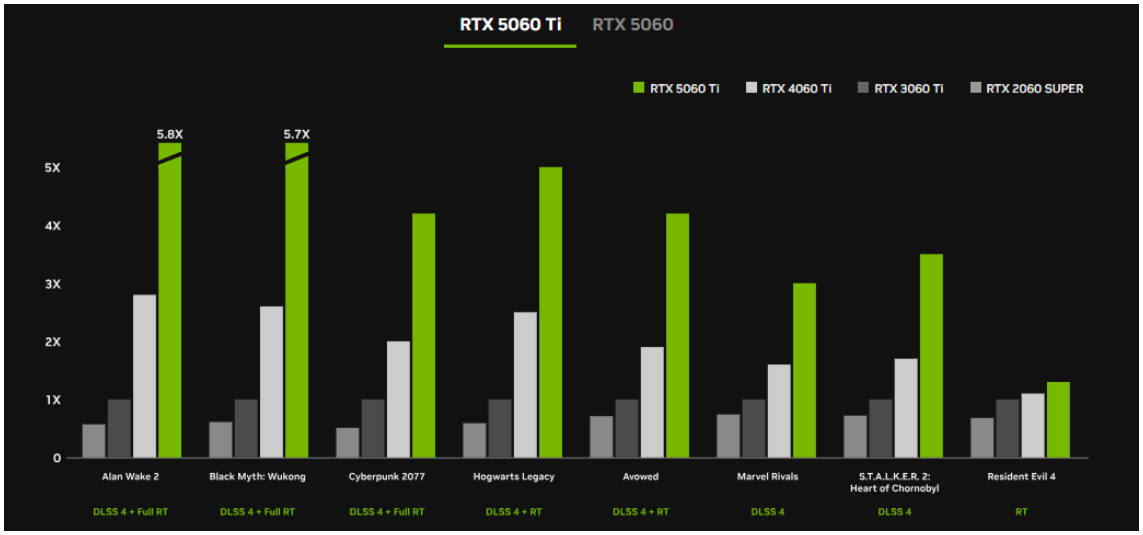
Este trabalho adota uma abordagem qualitativa, de caráter exploratório e descritivo, fundamentada em uma revisão de literatura e na análise de exemplos práticos do mercado.

A pesquisa bibliográfica foi realizada por meio da análise de livros, artigos científicos, periódicos acadêmicos e materiais técnicos. As fontes foram levantadas nas bases IEEE Xplore, Scopus, Google Scholar, além de publicações de empresas da área de tecnologia. Este levantamento teve como objetivo compreender os conceitos fundamentais da inteligência artificial aplicada ao *design* de *hardware*, as arquiteturas de processadores neurais e os principais desafios e soluções disponíveis.

Para ilustrar os conceitos levantados na literatura, o estudo apresenta dados de desempenho divulgados por fabricantes de *hardware*. Esses exemplos demonstram a aplicação prática das técnicas de otimização:

Caso NVIDIA: É apresentado um comparativo de desempenho da linha de placas de vídeo RTX 50, demonstrando a melhoria gráfica em jogos com o uso de tecnologias baseadas em IA (Figura 1).

Figura 1: Teste de desempenho de placas de vídeo




Fonte: NVIDIA (2025).

Caso Intel: É apresentado um demonstrativo de desempenho da linha de processadores Intel Core Ultra (Quadro 1), que aponta o ganho em processamento neural e a redução no consumo energético em relação à geração anterior.

Quadro 1: Demonstrativo de comparação de processador INTEL Core Ultra Series

| Métrica | Geração Anterior | Core Ultra Series 2 |
|---|---|---|
| Desempenho em Data Science | 100% | 122% (≈ +22 %) |
| Desempenho Multithread | 100% | 119 % (≈ +19 %) |
| Consumo de Energia (média) | 100% | 56 % (≈ -44 %) |

Fonte: INTEL, adaptado (2024)

## 4 RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados obtidos através da revisão de literatura e da análise dos exemplos práticos, conforme descrito na metodologia. Os achados são divididos em resultados teóricos (da literatura) e práticos (da indústria), seguidos de uma discussão que os conecta à problemática central do trabalho.

### 4.1 Achados da Revisão de Literatura

A revisão da literatura permitiu identificar as principais frentes de otimização de *hardware* usando IA, que servem de base para os avanços atuais:

Validação da Eficiência de NPUs: A literatura confirma a premissa central do estudo. Sze *et al*. (2017) demonstram que arquiteturas tradicionais (CPUs/GPUs) não são ideais para IA, e que os Núcleos de Processamento Neural (NPUs) oferecem melhorias significativas em eficiência energética, latência e custo.

IA no Processo de *Design* (Aprendizado por Reforço): O principal achado é o uso da IA para otimizar o próprio *design*. O trabalho de Mirhoseini (2021) é um resultado fundamental, provando que o aprendizado por reforço pode superar engenheiros humanos no posicionamento de blocos de circuitos, reduzindo o tempo de desenvolvimento de semanas para horas e criando *designs* mais eficientes.

IA em Otimização Multiobjetivo (Algoritmos Genéticos): A pesquisa identificou que Algoritmos Genéticos (GAs) são uma técnica robusta para resolver problemas complexos de *design* de *hardware*, especialmente em otimização multiobjetivo, onde é preciso balancear métricas concorrentes como desempenho, custo e consumo energético (DEB *et al*., 2002).

## 4.2 Análise dos Exemplos Práticos

Os exemplos de mercado analisados na metodologia demonstram a aplicação prática dos conceitos teóricos:

Caso NVIDIA (Desempenho em IA): Os dados da NVIDIA (Figura 1) mostram o resultado prático da integração de hardware especializado (núcleos de IA). A RTX 5060 Ti, ao usar IA para processamento gráfico (DLSS 4), atinge um desempenho até 5.8x superior à geração RTX 2060 em tarefas idênticas. Isso valida o impacto dos núcleos neurais dedicados no processamento de IA.

Caso Intel (Eficiência e Custo): Os dados da Intel (Quadro 1) abordam diretamente os objetivos centrais deste TCC: melhoria de processamento e redução de custos (energéticos). A nova geração "Core Ultra Series" apresenta um ganho de 22% em *Data Science* (processamento) e, crucialmente, uma redução de 44% no consumo de energia. Isso corrobora a tese de que a otimização via IA leva a *hardware* mais eficiente e sustentável.

## 4.3 Discussão

Os resultados, tanto da literatura quanto da indústria, convergem e respondem à pergunta de pesquisa. A crescente demanda por IA impõe desafios significativos de *design*, custo e eficiência energética.

A discussão destes resultados evidencia que as técnicas de IA não são apenas o objetivo do novo *hardware*, mas também o meio para o criar. O trabalho de Mirhoseini (2021) é a prova teórica de que a complexidade do *design* pode ser superada pela automação inteligente. Por sua vez, os dados da Intel são a prova de mercado de que essa otimização resulta em produtos com drástica redução de consumo energético, atacando o desafio do custo operacional e da sustentabilidade.

Portanto, a aplicação de técnicas de inteligência artificial no próprio processo de desenvolvimento de *hardware* é a estratégia que permite a automação, redução de erros e otimização de recursos que a indústria necessita. O estudo demonstra que a IA é um elemento ativo na concepção de *hardware*, e a relevância disso se intensifica com a expansão dos processadores neurais em todos os setores. A democratização do acesso a tecnologias de alto desempenho, mencionada como objetivo, passa diretamente pela capacidade de usar IA para baratear e tornar mais eficiente a produção do próprio *hardware* que a executa.

## 5 CONCLUSÃO

O presente trabalho teve como objetivo investigar o uso de técnicas de Inteligência Artificial (IA) na otimização de arquiteturas de *hardware*, com ênfase nos processadores que incorporam núcleos neurais dedicados. Por meio de uma abordagem qualitativa, exploratória e descritiva, baseada em revisão bibliográfica e análise de casos reais, foi possível compreender de que forma algoritmos de aprendizado de máquina, aprendizado por reforço e algoritmos genéticos têm contribuído para aprimorar o desempenho, a eficiência energética e a redução dos custos de fabricação de *hardware* especializado.

Os resultados obtidos evidenciam que a integração entre IA e *design* de *hardware* representa uma transformação significativa na indústria de tecnologia, permitindo que os próprios sistemas inteligentes participem do processo de criação e otimização de componentes. Essa evolução reflete-se em produtos mais eficientes e sustentáveis, como exemplificado pelas linhas mais recentes de processadores e placas de vídeo das empresas Intel e NVIDIA, que incorporam NPUs e núcleos de IA voltados à aceleração de tarefas de aprendizado de máquina e processamento neural.

Constatou-se também que o uso da IA na otimização de *hardware* possibilita maior automação no ciclo de desenvolvimento, redução de erros e melhor aproveitamento dos recursos físicos e energéticos. Tais avanços contribuem diretamente para a democratização do acesso a tecnologias de alto desempenho, beneficiando setores como saúde, indústria, educação e segurança digital.

Por fim, conclui-se que a aplicação da Inteligência Artificial no design e otimização de hardware não apenas promove ganhos técnicos, mas também inaugura uma nova era de

integração entre *software* e *hardware* inteligente. Como proposta de continuidade, recomenda-se a realização de estudos experimentais com protótipos de processadores neurais e medições práticas de desempenho, a fim de validar empiricamente os benefícios apresentados na literatura e fortalecer o desenvolvimento de soluções computacionais mais acessíveis e sustentáveis.

## REFERÊNCIAS


DEB, K.; PRATAP, A.; AGARWAL, S.; MEYARIVAN, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, v. 6, n. 2, p. 182-197, 2002. DOI: https://ieeexplore.ieee.org/document/996017
GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge: MIT Press, 2016.

HAN, Song; MAO, Huizi; DALLY, William J. DEEP COMPRESSION: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. arXiv, 2024 [Original 2015]. Disponível em: https://ar5iv.labs.arxiv.org/html/1510.00149. Acesso em: 12/08/2025.

HARDWARE.COM.BR. O que são aceleradores de IA? Hardware.com.br, 2025. Disponível em: https://www.hardware.com.br/artigos/o-que-sao-aceleradores-de-ia/. Acesso em: 08/10/2025.

MIRHOSEINI, A. et al. A graph placement methodology using deep reinforcement learning. Nature, v. 593, p. 207–212, 2021. DOI: https://www.nature.com/articles/s41586-021-03544-w.

NVIDIA. Família GeForce, RTX 5060: Transformando o Jogo. [S.l.]: NVIDIA, 15 abr. 2025. Disponível em: https://www.nvidia.com/pt-br/geforce/graphics-cards/50-series/rtx-5060-family. Acesso em: 15/07/2025.

NVIDIA AND INTEL TO DEVELOP AI INFRASTRUCTURE AND PERSONAL COMPUTING PRODUCTS. NVIDIA Newsroom, 18 set. 2025. Disponível em: https://nvidianews.nvidia.com/news/nvidia-and-intel-to-develop-ai-infrastructure-and-personal-computing-products. Acesso em: 15/10/2025.

RUSSELL, Stuart; NORVIG, Peter. Inteligência Artificial. 4. ed. São Paulo: Pearson, 2021.

SZE, Vivienne; CHEN, Yu-Hsin; YANG, Tien-Ju; EMER, Joel S. Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, v. 105, n. 12, p. 2295-2329, 2017. DOI: https://ieeexplore.ieee.org/document/8114708.