Jornada de Iniciação Científica

SEMANA DE INTEGRAÇÃO ACADÊMICA

Aprendizagem Profunda e IA Verde:

Caminhos Para Um Futuro Mais Sustentável

UNIRIO
UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO

PROPGPI
PRÓ-REITORIA DE PÓS-GRADUAÇÃO, PESQUISA E INOVAÇÃO

DPQ
DIRETORIA DE PESQUISA

Discente: Vívian Rique Gil Ferraro

Professor Responsável: Pedro Nuno de Souza Moura

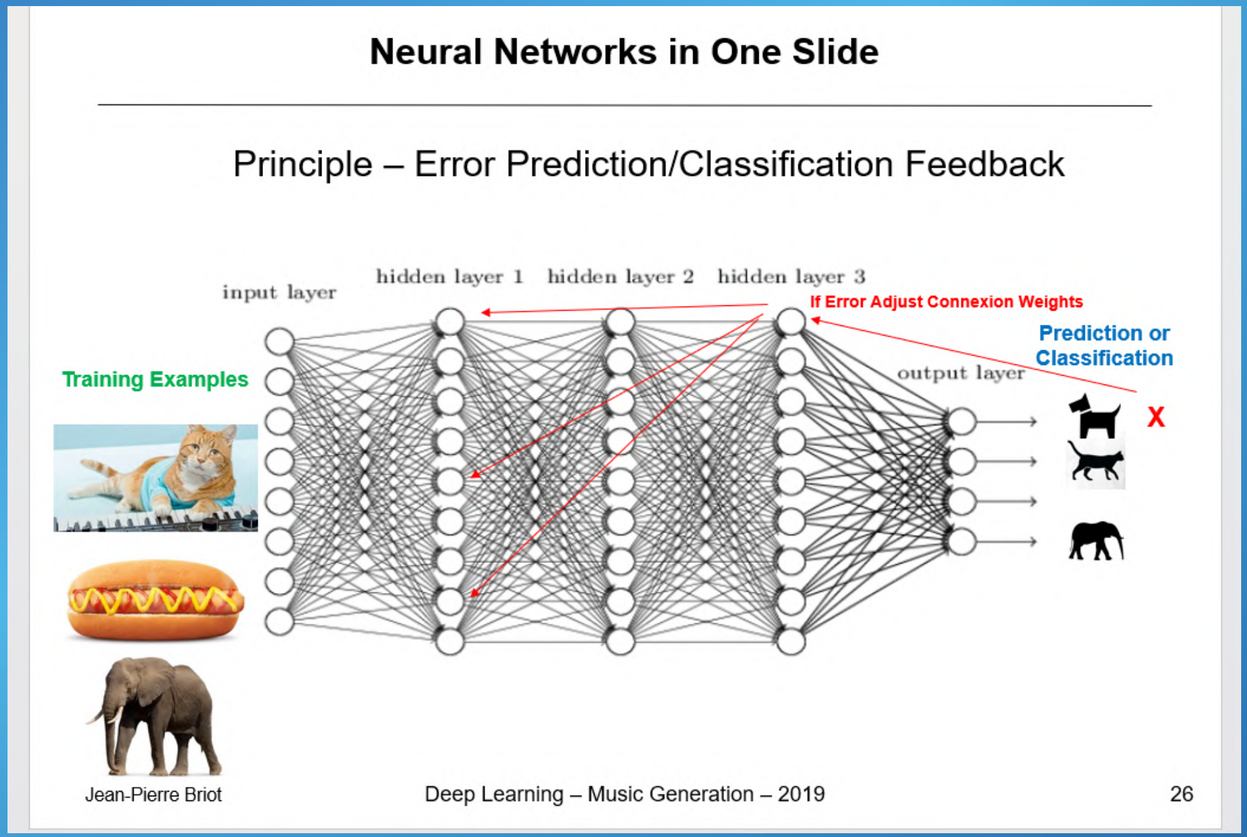
Coorientação: Daniel da Silva Costa

Aprendizagem Profunda

A aprendizagem profunda é um subcampo da inteligência artificial que se concentra na construção e treinamento de redes neurais.

- Modelos capazes de realizar tarefas complexas com grande precisão.

Aprendizagem Profunda



Qual o preço?

Modelos cada vez mais robustos, com grande precisão e capacidade de processamento.

Aumento do consumo de energia e dos recursos computacionais.

Custos cada vez maiores

01
BERT-large

- 350 milhões de parâmetros

02
GPT2-XL

- 1,5 bilhão de parâmetros

03
GPT3

- 175 bilhões de parâmetros

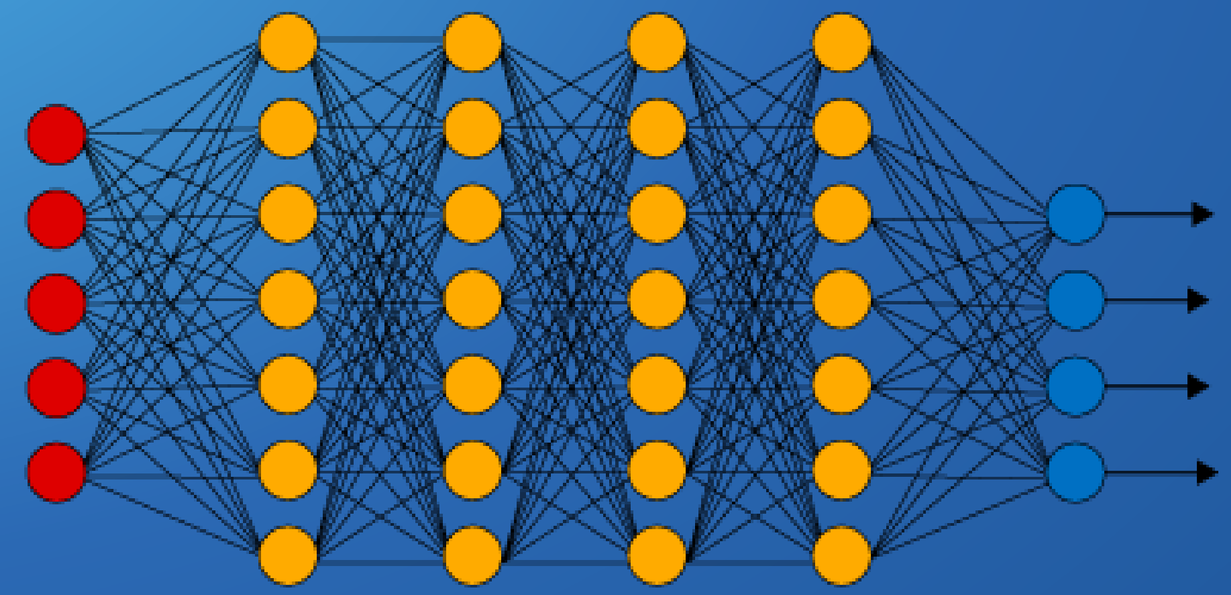
04
GPT4

- superior a 1 trilhão de parâmetros

O que são parâmetros?

São valores numéricos que a rede usa para tomar decisões.

Funcionam como ‘regras’ internas da rede, que permitem uma compreensão mais profunda dos dados e a capacidade de fazer previsões mais precisas.



https://www.deeplearningbook.com.br/o-que-sao-redes-neurais-artificiais-profundas/

Experimentos

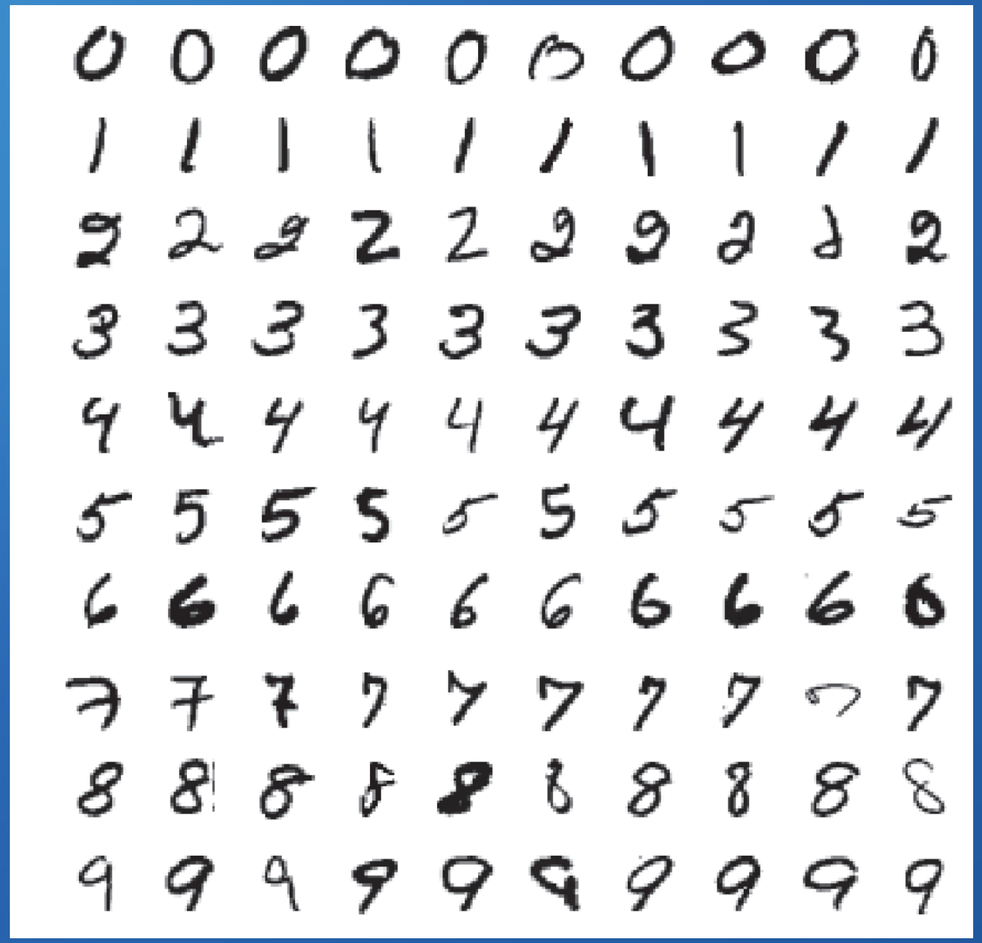
Base de dados:

MNIST

(LECUN, 1998)

Conjunto de treinamento com **60.000** exemplos

Conjunto de testes com **10.000** exemplos

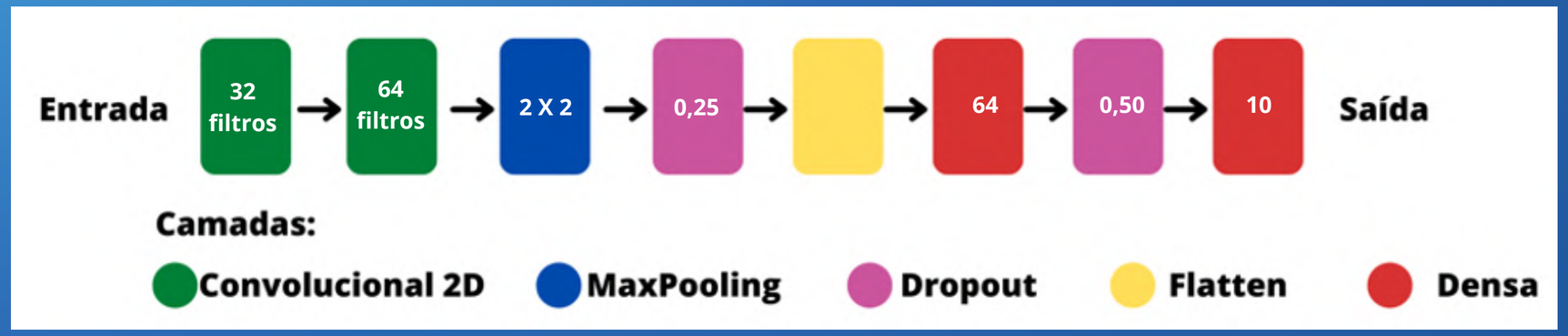


Experimentos

Rede utilizada:

LeNet-5

(LECUN et al., 1998)



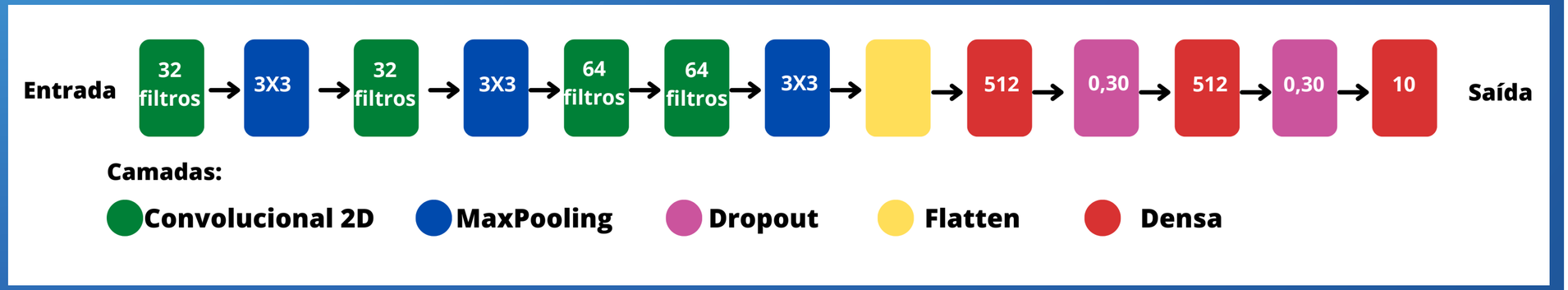
Número de parâmetros: 609.354

Experimentos

Rede utilizada:

AlexNet

(KRIZHEVSKY et al., 2017)



Número de parâmetros: 1.938.922

Métricas adotadas

Gasto Energético

- Tempo de treino
- Tempo de inferência
- Número de parâmetros
- Operações de ponto flutuante
- Consumo energético do Hardware

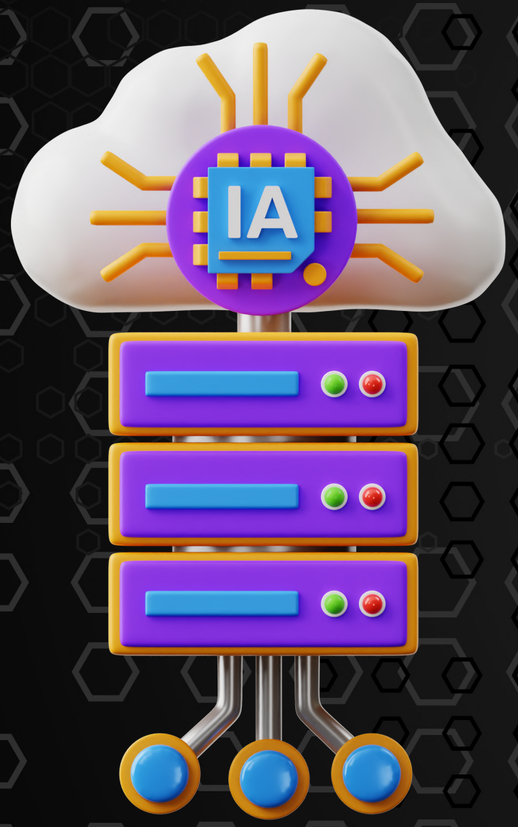
Desempenho

- Acurácia
- Precisão
- Recall
- F1-Score

Desenvolvimento

- Foram realizados 10 treinos, que geraram 10 modelos para cada rede.
- O modelo que obteve maior acurácia em cada rede foi escolhido para a inferência.
- O conjunto de dados de treino foi dividido em treino (80%) e validação (20%).
- O tamanho do batch foi de 128.
- Cada treino foi realizado em 20 épocas.

Características do ambiente de treino e testes



- Plataforma Google Colab
- GPU NVidia V-100
- 12,68 GB memória RAM

Principais recursos utilizados

- Linguagem Python
- Interface Jupyter Notebook
- Biblioteca TensorFlow
- Biblioteca Keras_flops para medir o número de FLOPS
- Biblioteca pyJoules para medir o consumo energético

Resultados

Métricas de consumo para cada rede

Modelos	Tempo de treino (segundos)	Tempo de inferência (milissegundos)	Operações de ponto flutuante (GFLOPS)	Consumo Energético (microjoules)	Número de Parâmetros
LeNet-5	33,8262	854	2,93	3956288	609.354
AlexNet	76,645768	1737,982	5,67	6889739,4	1.938.922

*Os valores para tempo de treino e consumo energético foram obtidos durante os treinos, através das média dos valores de cada modelo gerado pelas redes.

*O tempo de inferência foi obtido no teste como o modelo com maior acurácia de cada rede.

*O número de operações de ponto flutuante é constante para cada rede.

Unidades:

1 GFLOP = 1×10^{9} FLOPS

1 milissegundo = 1×10^{-3} segundos

1 microJoule = 1×10^{-6} Joules

Resultados

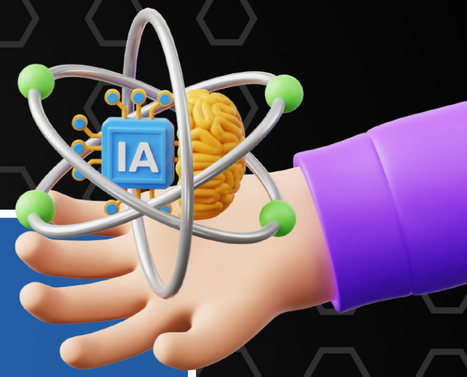
Desempenho de cada rede

Modelo	Acurácia	Precisão	Recall	F1-Score
LeNet-5	0,9929	0,992841	0,992858	0,992845
AlexNet	0,985	0,984997	0,984776	0,984865

*Valores obtidos durante a inferência.

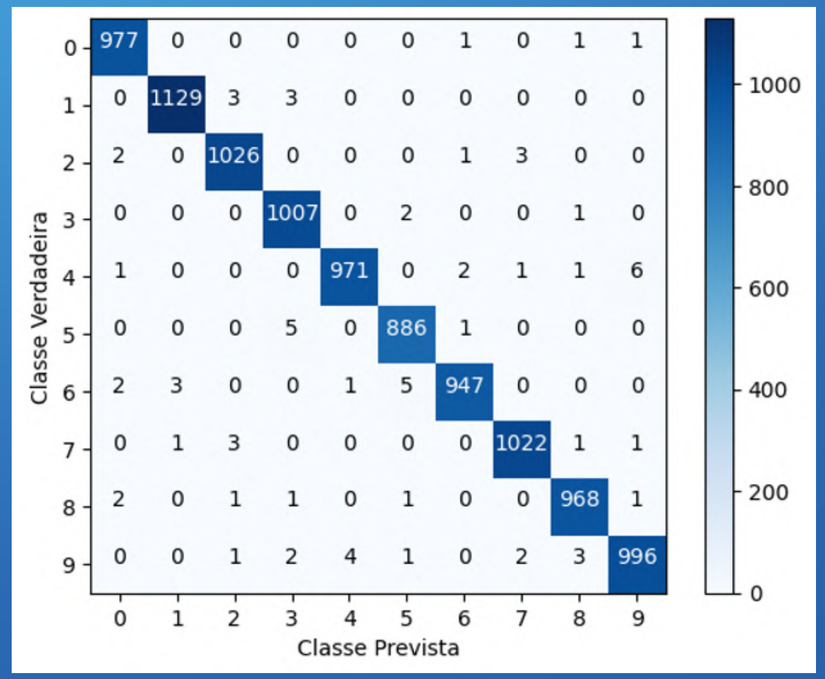
* Valores normalizados

Resultados

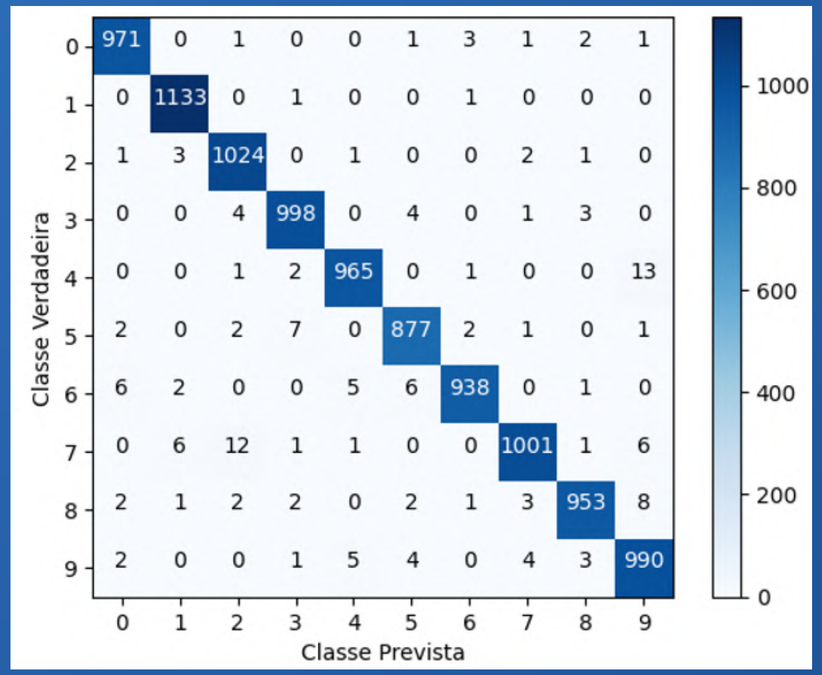


Matriz de confusão

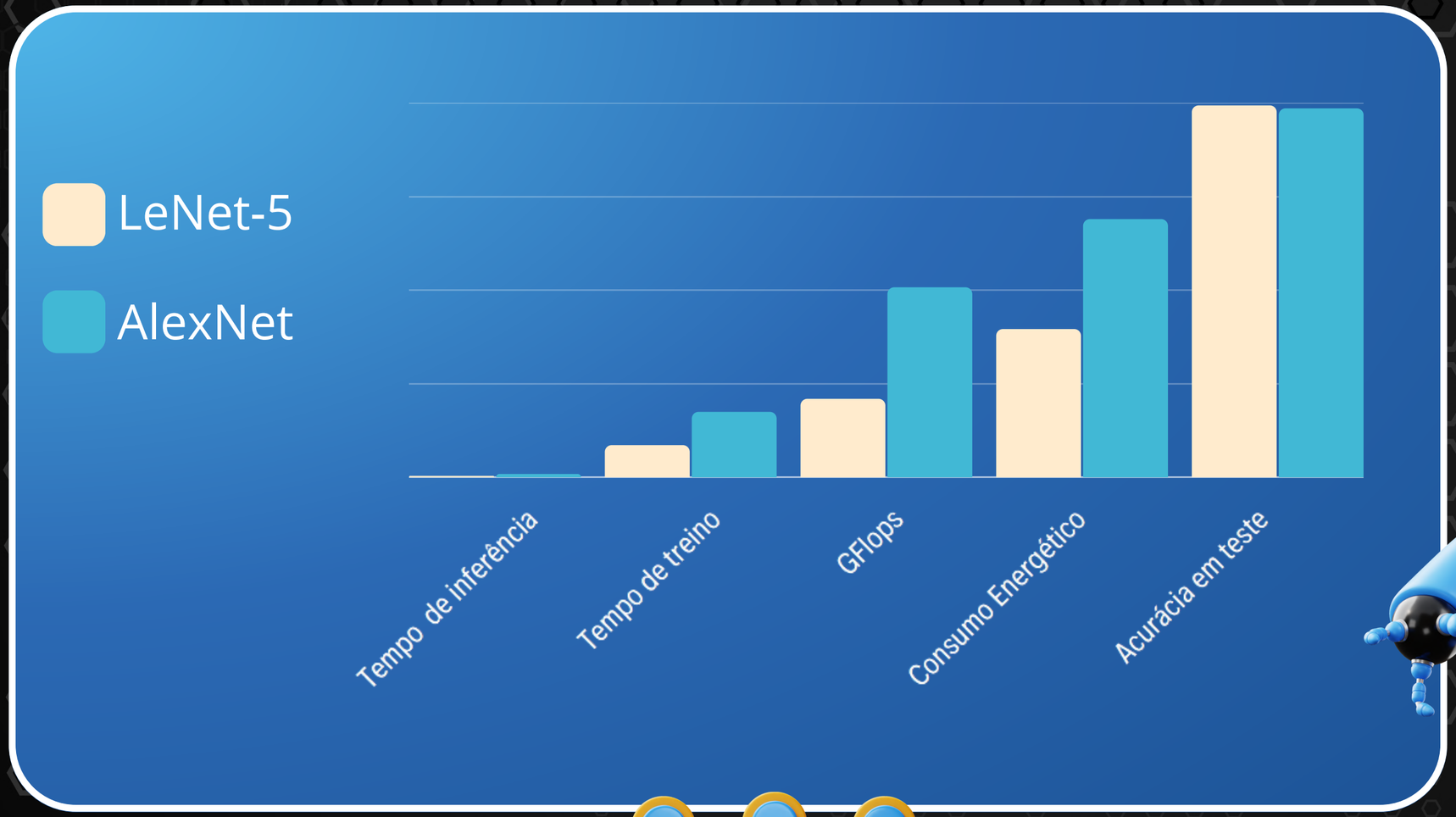
LeNet-5



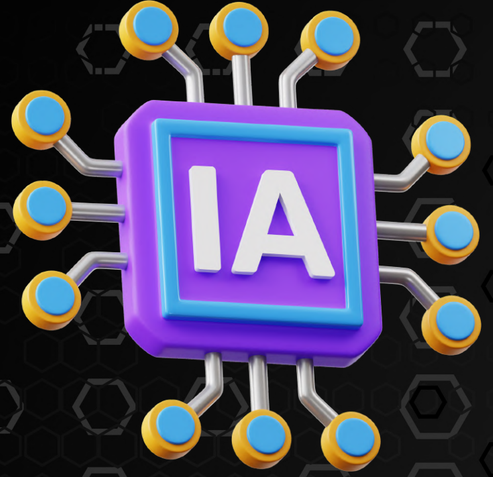
AlexNet



Resultados


LeNet-5
AlexNet
Tempo de inferência
Tempo de treino
GFlops
Consumo Energético
Acurácia em teste

IA



Conclusões

- Os resultados serviram para validar as métricas levantadas na literatura.
- Conforme esperado, os modelos gerados com a rede AlexNet apresentaram valores maiores nas métricas de consumo energético adotadas.
- Apesar disso, o modelo produzido pela rede AlexNet não apresentou o melhor desempenho nos testes realizados.
- Ressaltando a importância de se considerar não apenas o desempenho dos modelos.

“A IA verde é uma oportunidade de criar um futuro mais sustentável para todos. Ao desenvolver e usar sistemas de IA de forma responsável, podemos reduzir nosso impacto ambiental e proteger o planeta para as gerações futuras.”

Marc Benioff,
cofundador e CEO da Salesforce

Referências

- ANTHONY, Lasse F. Wolff; KANDING, Benjamin; SELVAN, Raghavendra. Carbontracker: tracking and predicting the carbon footprint of training deep learning models. arxiv preprint, arXiv:2007.03051, 2020.
- DESISLAVOV, Radosvet; MARTÍNEZ-PLUMED, Fernando; HERNÁNDEZ-ORALLO, José. Trends in AI inference energy consumption: beyond the performance-vs-parameter laws of deep learning. Sustainable Computing: Informatics and Systems, [S.L.], v. 38, p. 100857, abr. 2023.
- DOUWES, Constance; ESLING, Philippe; BRIOT, Jean-Pierre. Energy consumption of deep generative audio models. arXiv preprint, arXiv:2107.02621, 2021.
- INRIA - Institut National de Recherche en Informatique et en Automatique. pyJoules 0.2.0 documentation. Disponível em: <https://pyjoules.readthedocs.io/en/latest/>. Acesso em: 22 ago. 2023.
- KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey. ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM, [S.L.], v. 60, n. 6, p. 84–90, 2017.
- KROHN, Jon; BEYLEVELD, Grant; BASSENS, Aglaé. Deep Learning Illustrated: a visual, interactive guide to artificial intelligence. Boston: Pearson Addison-Wesley, 2020. 416 p.
- LACOSTE, Alexandre; LUCCIONI, Alexandra; SCHMIDT, Victor; DANDRES, Thomas. Quantifying the carbon emissions of machine learning. arXiv preprint, arXiv:1910.09700, 2019.
- LECUN, Yann et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, v. 86, n. 11, p. 2278-2324, 1998.
- LENHERR, Nicola; PAWLITZEK, René; MICHEL, Bruno. New universal sustainability metrics to assess edge intelligence. Sustainable Computing: Informatics and Systems, [S.L.], v. 31, p. 100580, set. 2021.
- SCHWARTZ, Roy; DODGE, Jesse; SMITH, Noah A.; ETZIONI, Oren. Green AI. Communications of the ACM, [S.L.], v. 63, n. 12, p. 54-63, 2020.

Obrigada pela atenção!

Repositório do projeto no GitHub

viferraro@edu.unirio.br

Apresentação baseada no design ‘Blue and Black Futuristc Illustrative Artificial Intelligence project Presentation’, de Olmos Carlos e disponibilizado em https://www.canva.com.