

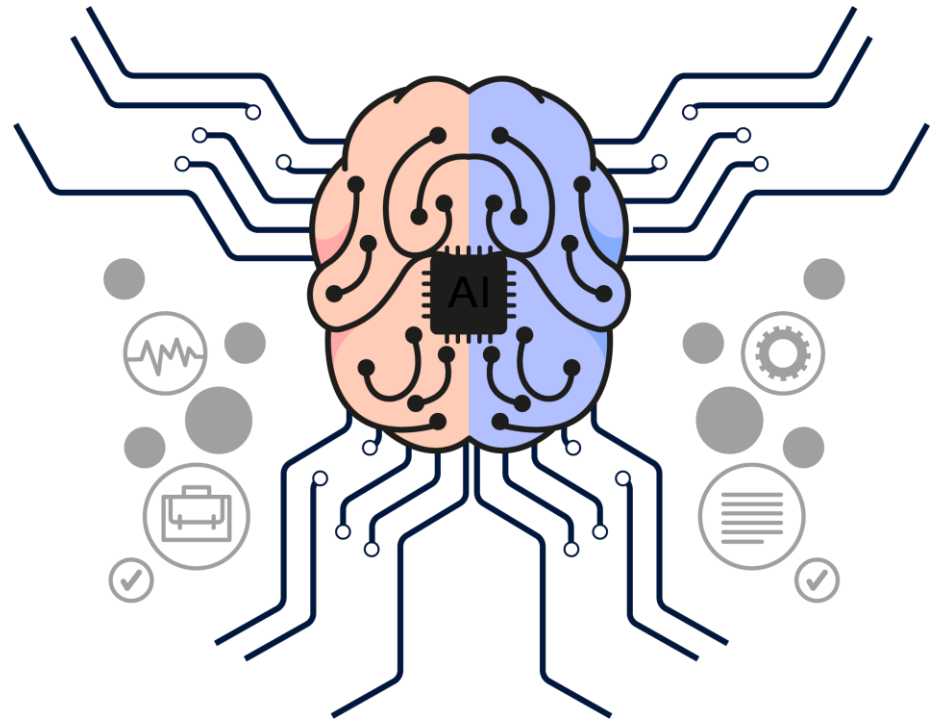


Paralelismo en IA

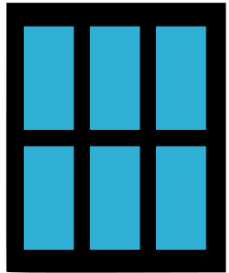
Axel Rodrigo Sotelo Ramírez



¿Qué es la IA?

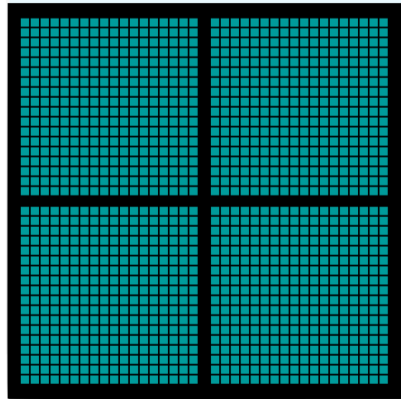


¿Procesador?
¿Tensor?



CPU
Multiple Cores

+





GPU
Thousands of Cores

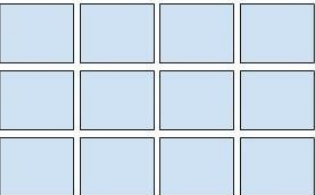
GPU vs CPU

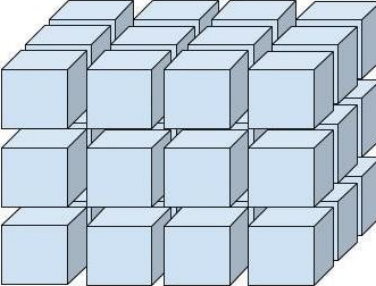
Tensores

¿Multiplicación de Tensores?

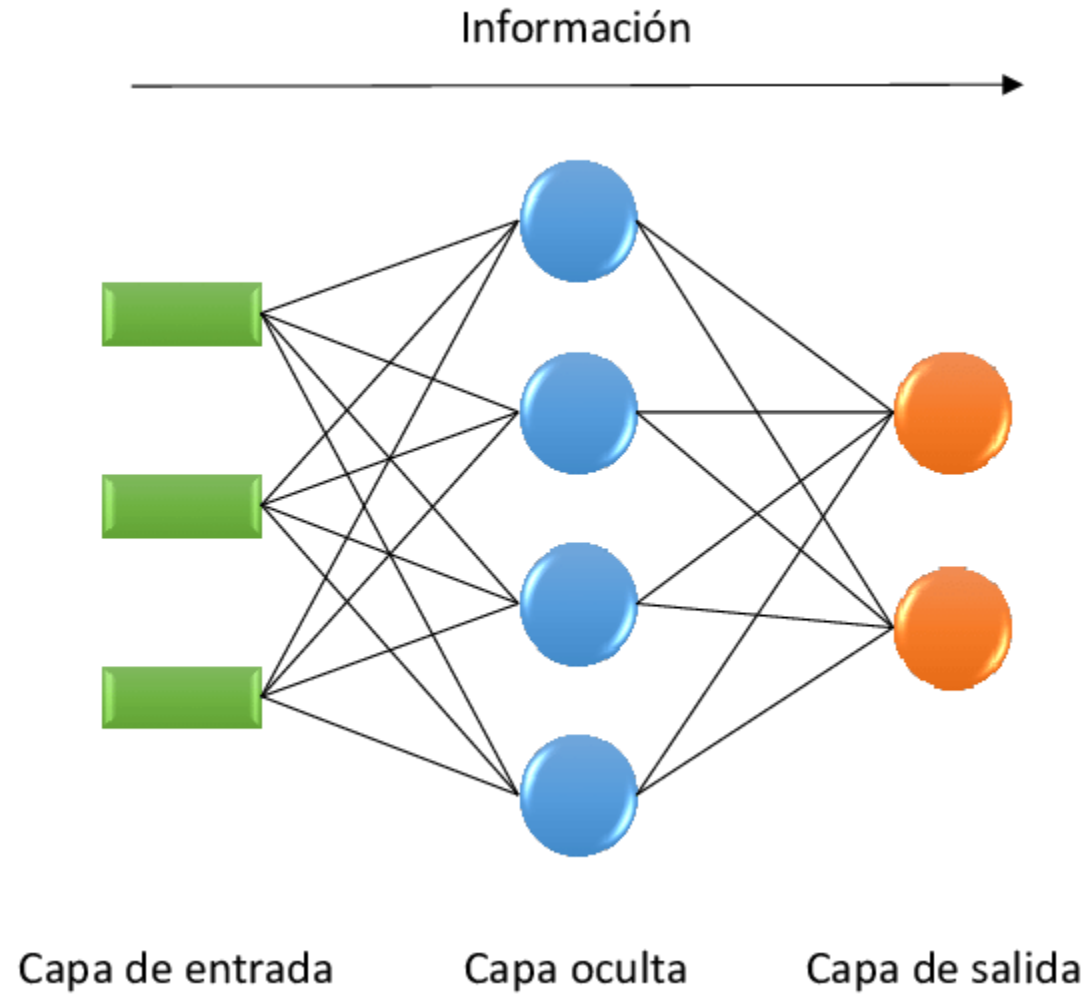
Rank 0: 
(scalar)

Rank 1: 
(vector)

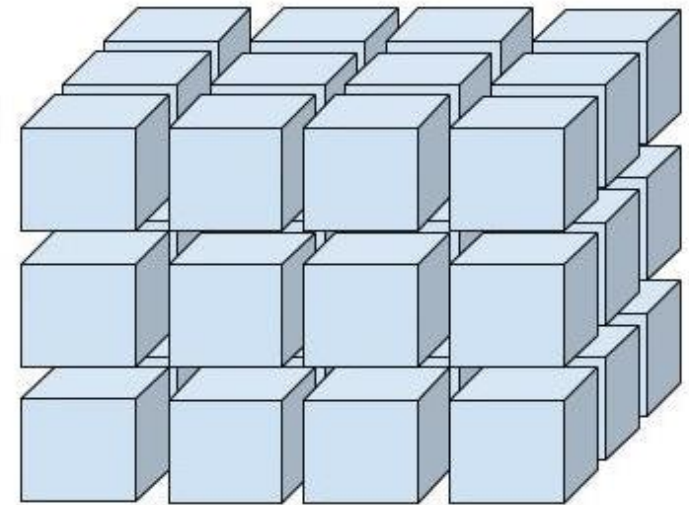
Rank 2: (matrix)


Rank 3: 

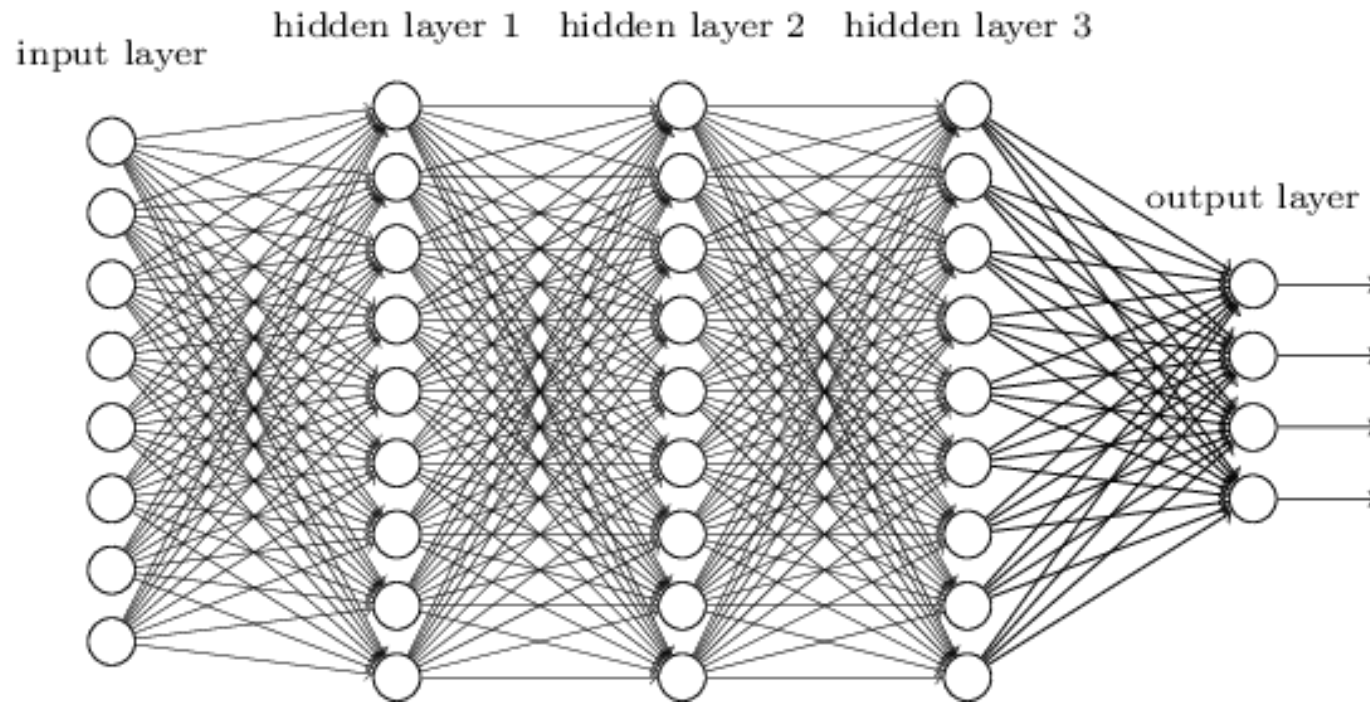
Red neuronal



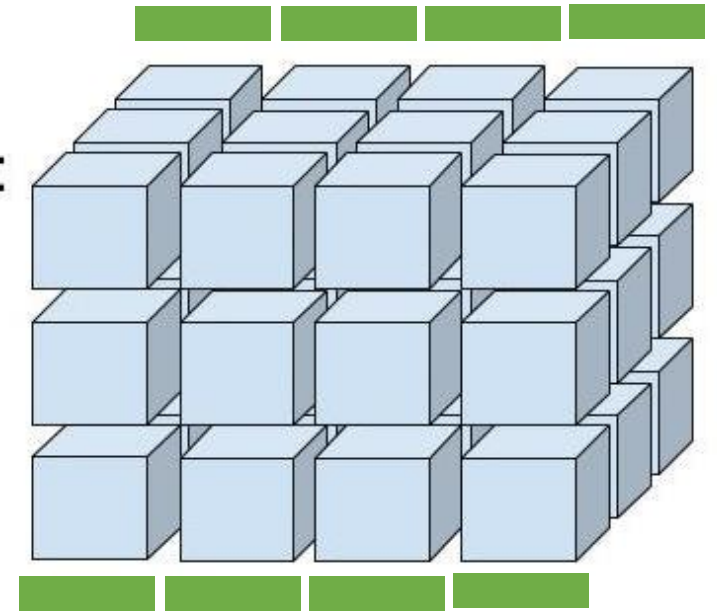
Rank 3:



Red neuronal



Rank 3:





Paralelismo

Tarjetas gráficas NVIDIA

Arquitectura

Arquitectura Fermi: Lanzada en 2010

Arquitectura Kepler: Lanzada en 2012

Arquitectura Maxwell: Lanzada en 2014

Arquitectura Pascal: Lanzada en 2016

Arquitectura Volta: Lanzada en 2017

Arquitectura Turing: Lanzada en 2018

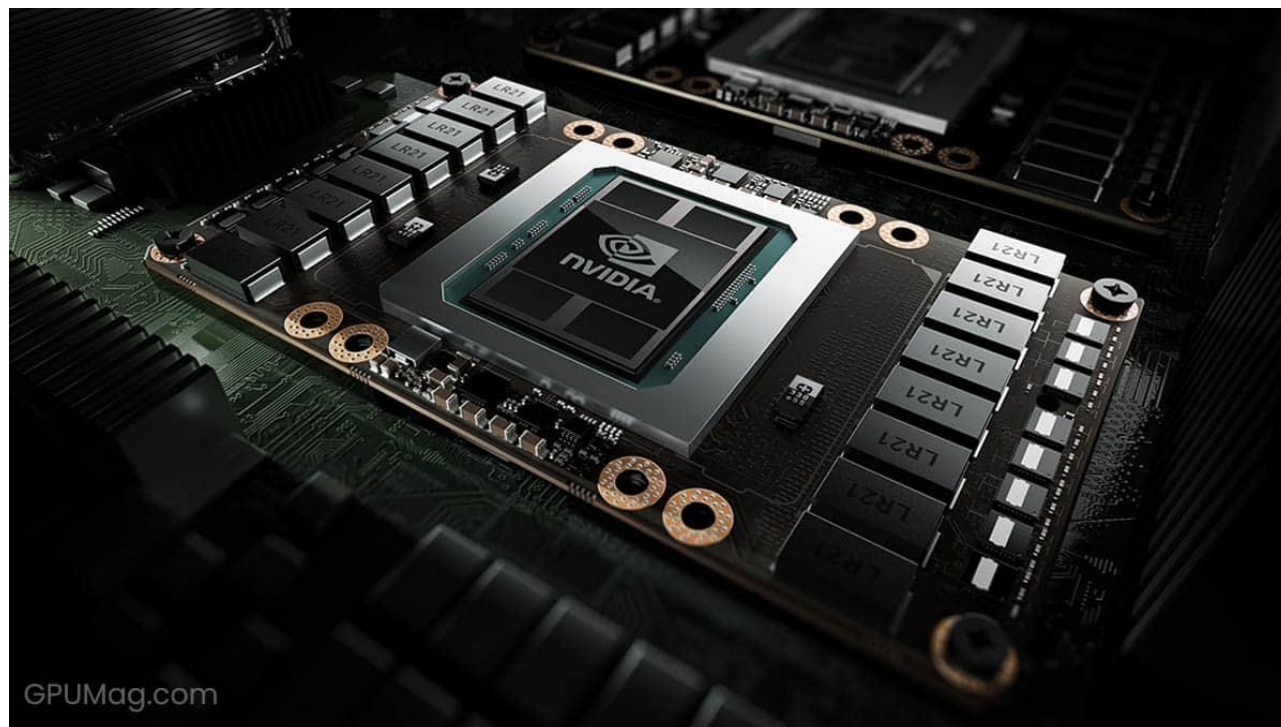
Arquitectura Ampere: Lanzada en 2020

CUDA Cores

TENSOR Cores

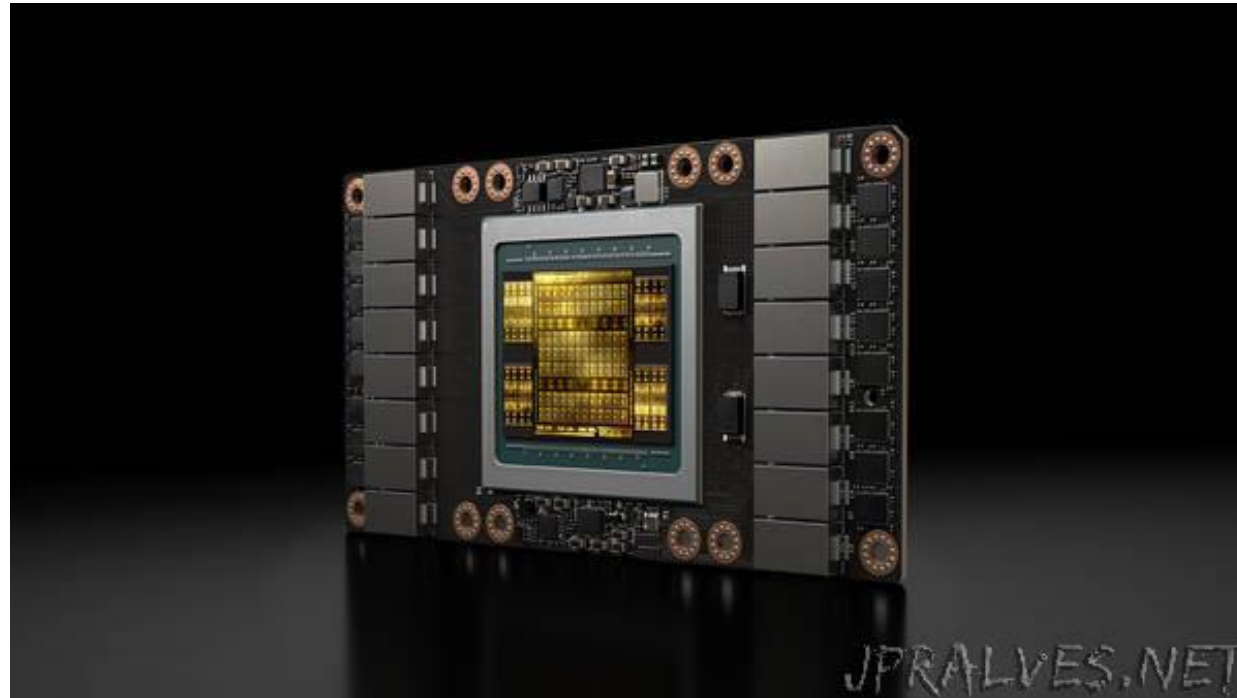


CUDA Cores



NVIDIA Tensor core

¿Qué es?



[Esta foto](#) de Autor desconocido está bajo licencia [CC BY-NC-ND](#)

NVIDIA Tensor core

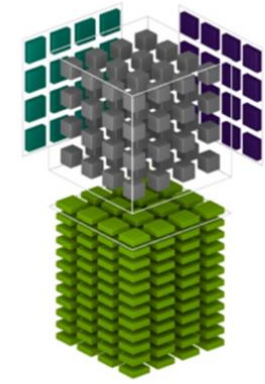
$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

HMMA FP16 or FP32
IMMA INT32

FP16
INT8 or UINT8

FP16
INT8 or UINT8

FP16 or FP32
INT32



Referencias

- [1]: Investopedia. (2022). Artificial Intelligence (AI). Retrieved May 16, 2023, from <https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp>
- [2]: Procesador - Concepto, componentes y funcionamiento. (2021). Concepto.de. <https://concepto.de/procesador/>
- [2]: ¿Qué es un Procesador y para qué sirve? - Definición. (2022). GEEKNETIC. <https://www.geeknetic.es/Procesador/que-es-y-para-que-sirve>
- [3]: ¿Qué es el núcleo de un procesador? - Profesional Review. (2021, febrero 13). Recuperado 16 de mayo de 2023, de <https://www.profesionalreview.com/2021/02/13/nucleo-procesador/>
- [4]: Deep Learning para todos los públicos: ¿Qué son los tensores? ¿Qué es TensorFlow? Think Big Empresas. Recuperado de <https://empresas.blogthinkbig.com/deep-learning-para-todos-los-publicos/>
- [5] ¿Qué es una red neuronal? - MATLAB & Simulink – MathWorks
- [6]: - Tipos de paralelismo computacional | KeepCoding Bootcamps. (s.f.). Recuperado el 14 de mayo de 2023, de <https://keepcoding.io/blog/tipos-de-paralelismo-computacional/>
- [7]: : NVIDIA (s.f.). Tecnologías y arquitecturas de GPU de NVIDIA | NVIDIA. Recuperado el 16 de mayo de 2023, de <https://www.nvidia.com/es-es/technologies/>
- [8]: HardZone. (2022). NVIDIA CUDA Cores: ¿Qué son y para qué sirven? [Blog post]. Retrieved May 16, 2023, from <https://hardzone.es/marcas/nvidia/nucleos-cuda/>
- [9]: NVIDIA Developer. (s.f.). Tensor Cores. Recuperado de <https://developer.nvidia.com/tensor-cores>
- [9] NVIDIA. (s.f.). Tensor Cores: versatilidad para HPC e IA. Recuperado de <https://www.nvidia.com/es-es/data-center/tensor-cores/>



Paralelismo en IA

Axel Rodrigo Sotelo Ramírez

