

## Paralelismo en Inteligencia Artificial

La inteligencia artificial ha tomado gran relevancia durante los últimos meses, por lo que hay herramientas que por lo mismo se han hecho demasiado populares, tal es el caso de ChatGPT, Midjourney, etc. Resulta que hoy en día es muy sencillo comenzar a programar un algoritmo de aprendizaje automático, ya que solo hace falta instalar un lenguaje de programación como Python y una que otra biblioteca para desarrollo de IA como Tensorflow, pytorch, etc. Pero ¿Cómo es que el sistema operativo se comunica con los componentes de la computadora para que nuestra IA se entrene efectivamente? Bueno pues para conocer y poder adentrarnos más en este tema, debemos de conocer antes algunos conceptos clave.

Ya conocemos lo que es un procesador, sabemos cómo funciona y su importancia en el área de la computación, pero hay algo en lo que nos debemos de fijar al momento de que comparamos un CPU con una GPU, hablo de sus respectivos núcleos. Resulta que los núcleos que tiene un CPU son de propósito general, es decir, sirven para procesar una gran cantidad de tareas de distinta índole, es por eso que un CPU suele tener entre 2, 4, 8 e incluso 32 núcleos, pero la GPU llega a tener hasta 2000 núcleos, ¿cuál es la diferencia? Ésta última radica en que los núcleos de la GPU son de propósito específico, lo que quiere decir que contienen una serie de instrucciones especializadas en un área específica de la computación, que por lo general, es el área gráfica.

Ya que conocemos la principal diferencia relevante para nuestro pequeño artículo, vamos a definir lo que es un tensor. Un tensor es aquella estructura de datos que se puede expresar o representar como un arreglo de matrices. El tensor es una estructura de datos muy relevante para el estudio y desarrollo de la IA ya que nos permite representar a las redes neuronales de mejor manera.

Ya que tocamos el tema de los tensores podemos ya empezar a analizarlos, y ver que hacer una multiplicación de tensores es, computacionalmente hablando, muy complicado, ya que supone tener un ciclo dentro de otro ciclo, dentro de otro ciclo para poder recorrer el número de tensores que se quieran multiplicar. Esto en programación secuencial es muy difícil de lograr, pero se resuelve si hablamos de programación paralela.

La programación paralela la hemos hablado mucho durante el desarrollo del curso de sistemas operativos del profesor, por lo que ya tenemos una idea clara de a lo que se refiere. Sin embargo, vamos a ver cómo hace NVIDIA para poder utilizar el paralelismo a su favor a través de sus núcleos.

Las gráficas NVIDIA se pueden clasificar en su arquitectura, en este caso hablaremos de la arquitectura fermi y la volta, ya que fue en la arquitectura fermi cuando se anunció por primera vez el concepto de CUDA core, y fue en la arquitectura Volta cuando se anunció el concepto de Tensor core.

Pero ¿qué es un CUDA core? Imagina que un CUDA core es como un pequeño cerebro especializado en hacer cálculos muy rápidos. Lo podemos ver como un superhéroe matemático que puede hacer esos cálculos en un abrir y cerrar de ojos.

Cuando usas una tarjeta gráfica o GPU para jugar videojuegos o hacer cosas intensivas en gráficos, los CUDA cores son los encargados de procesar toda esa información visual de manera rápida y eficiente. Son como pequeños trabajadores que se dividen el trabajo y realizan cálculos simultáneamente para que todo se vea suave y sin demoras en la pantalla.

Imagina que estás jugando un juego súper exigente y necesitas que la tarjeta gráfica procese miles de imágenes por segundo para que la acción sea fluida. Los CUDA cores se ponen manos a la obra y hacen todo ese trabajo pesado para que puedas disfrutar de una experiencia de juego increíble.

¿Qué es un tensor core? Un Tensor Core es otro tipo de unidad especializada en las tarjetas gráficas o GPU. Podemos decir que es como un maestro de las matemáticas avanzadas y los cálculos intensivos. Cuando estás trabajando en un proyecto de IA o entrenando un modelo de aprendizaje automático, hay muchas operaciones matemáticas que deben realizarse repetidamente. Aquí es donde entran en acción los Tensor Cores, ellos pueden realizar cálculos vectoriales y matriciales de alta precisión de manera simultánea y ultrarrápida. Pero aquí está la magia: los Tensor Cores no solo son rápidos, sino que también pueden realizar operaciones matemáticas mixtas con baja precisión, lo que se conoce como precisión reducida. Esto significa que pueden hacer los cálculos necesarios con menos precisión, pero sin comprometer demasiado la calidad del resultado. Esto ahorra tiempo y energía, lo cual es increíble.

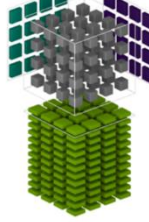
Estos núcleos están diseñados específicamente para realizar operaciones de multiplicación de matrices de manera eficiente y altamente paralela. Aprovechan la arquitectura optimizada de la GPU para llevar a cabo estos cálculos complejos en paralelo y en tiempo real.

Los Tensor Cores pueden procesar múltiples operaciones de multiplicación y suma en paralelo utilizando técnicas de precisión reducida, como la aritmética de punto flotante de baja precisión. Esto significa que pueden realizar cálculos matemáticos con menos precisión, pero sin comprometer significativamente la calidad de los resultados. Esta capacidad de realizar operaciones con precisión reducida permite un procesamiento más rápido y eficiente.

Al realizar operaciones de multiplicación de matrices con Tensor Cores, se logra un aumento significativo en el rendimiento y la velocidad de los cálculos utilizados en algoritmos de aprendizaje automático, como el entrenamiento de redes neuronales profundas. Estas unidades especializadas aceleran drásticamente el tiempo de entrenamiento de los modelos y permiten procesar grandes conjuntos de datos con mayor eficiencia.

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

HMMA FP16 or FP32      FP16      FP16      FP16 or FP32  
 IMMA INT32      INT8 or UINT8      INT8 or UINT8      INT32



### ¿Qué es inteligencia artificial?

se refiere a la capacidad de las máquinas para realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, la percepción visual, el reconocimiento del habla y la toma de decisiones. [1]

### ¿Qué es un procesador?

Un procesador es un circuito electrónico digital que ejecuta operaciones sobre un conjunto de datos. Se le considera el cerebro del sistema, ya que procesa todo lo que ocurre en la computadora y ejecuta todas las acciones que existen. Cuanto más rápido sea el procesador que tiene una computadora, más rápidamente se ejecutarán las órdenes que se le den a la máquina. El procesador es uno de los componentes de la computadora que más ha evolucionado, dado a que se les exige a los ingenieros que cada vez ofrezcan mejores procesadores para que las computadoras funcionen más rápidas y de forma más eficaz. [2]

### ¿Qué es un núcleo en la computadora?

Un núcleo en el procesador de una computadora es una unidad de procesamiento que lee las instrucciones y ejecuta acciones específicas. Cada núcleo es una unidad central de proceso separada e independiente. Los núcleos realizan cuatro tareas fundamentales: buscar, decodificar, ejecutar y reescribir. Cuantos más núcleos tenga un procesador, más tareas simultáneas se pueden realizar en la computadora. Los núcleos permiten conseguir un paralelismo en la ejecución que es imposible en un sistema mononúcleo. Además, cada núcleo físico se puede dividir en dos núcleos lógicos gracias a las tecnologías de multithreading. Los núcleos multiplican el rendimiento del procesador. El núcleo es la parte más importante del CPU, donde se procesa la información del sistema computacional. [3]

### ¿Qué es un Tensor?

Un tensor es un objeto matemático que almacena valores numéricos y que puede tener distintas dimensiones. En computación, los tensores se usan para representar y procesar datos de diferentes tipos, como números, imágenes, sonidos o textos. Los tensores se pueden clasificar según el número de dimensiones que tienen: un tensor de una dimensión

es un vector, un tensor de dos dimensiones es una matriz, un tensor de tres dimensiones es un cubo, y así sucesivamente. [4]

#### ¿Qué es una Red neuronal?

Una red neuronal en computación es un modelo de inteligencia artificial que imita el funcionamiento del cerebro humano, usando unidades llamadas neuronas artificiales que se conectan entre sí para procesar datos y aprender de ellos. Las redes neuronales pueden resolver problemas complejos y no lineales, como el reconocimiento de imágenes, el procesamiento de lenguaje natural o la predicción de eventos. Las redes neuronales se componen de capas de nodos, que incluyen una capa de entrada, una o más capas ocultas y una capa de salida. Cada nodo recibe datos de entrada, los multiplica por unos pesos, los suma y los pasa por una función de activación que determina la salida. Si la salida supera un umbral, el nodo se activa y envía datos a la siguiente capa. Las redes neuronales se entrenan con datos etiquetados para ajustar los pesos y mejorar su precisión con el tiempo. [5]

#### ¿Qué es el paralelismo?

El paralelismo en computación es una forma de realizar varios cálculos al mismo tiempo, aprovechando el uso de múltiples CPU o procesadores. El objetivo es resolver problemas grandes y complejos de manera más rápida y eficiente, dividiéndolos en subproblemas más pequeños que se ejecutan en paralelo. El paralelismo se puede aplicar a diferentes niveles, como el nivel de bit, el nivel de instrucción, el nivel de datos o el nivel de tarea. También se pueden clasificar los sistemas paralelos según el flujo de control o el flujo de datos que siguen. Algunas aplicaciones del paralelismo en computación son la bioinformática, la economía, la física y la inteligencia artificial. [6]

#### Arquitectura de las gráficas NVIDIA

Las tarjetas gráficas NVIDIA se basan en diferentes arquitecturas de GPU que combinan tecnologías como el ray tracing, la inteligencia artificial y el sombreado programable para ofrecer una experiencia de juego y de creación de contenido de alta calidad. Algunas de las arquitecturas más recientes son:

- Ampere: Es la arquitectura que utiliza la serie GeForce RTX 30, que ofrece un rendimiento superior al de la generación anterior gracias a sus núcleos CUDA mejorados, sus núcleos RT de segunda generación y sus núcleos Tensor de tercera generación. Además, cuenta con la memoria GDDR6X más rápida del mundo y soporta la tecnología NVIDIA DLSS, que mejora la calidad de imagen mediante el uso de redes neuronales.
- Turing: Es la arquitectura que utiliza la serie GeForce RTX 20 y la serie Quadro RTX, que fueron las primeras en introducir el ray tracing en tiempo real y la inteligencia artificial en los gráficos por computación. Turing cuenta con núcleos CUDA, núcleos RT de primera generación y núcleos Tensor de segunda generación, que permiten ejecutar algoritmos de IA en tiempo real para crear efectos visuales realistas y naturales.

- Pascal: Es la arquitectura que utiliza la serie GeForce GTX 10 y la serie Quadro P, que ofrecen un gran rendimiento y eficiencia energética para los juegos y las aplicaciones profesionales. Pascal cuenta con núcleos CUDA y soporta tecnologías como el sombreado simultáneo, la compresión de color delta y el sombreado anisotrópico mejorado. [7]

#### CUDA Core

Un CUDA core es un núcleo de procesamiento paralelo que se encuentra en las tarjetas gráficas de Nvidia y que se encarga de realizar cálculos gráficos. CUDA significa Compute Unified Device Architecture y es una plataforma de software que permite aprovechar la potencia de la GPU para realizar tareas de computación intensivas. Un CUDA core es similar a un núcleo de CPU, pero más pequeño y más numeroso. Los CUDA cores pueden renderizar los detalles de los gráficos digitales, como objetos 3D, iluminación, sombreado, etc., de forma simultánea y eficiente. [8]

#### Tensor Core

Un tensor core es una unidad especializada de procesamiento que se encuentra en algunas tarjetas gráficas de NVIDIA y que sirve para acelerar el entrenamiento y la inferencia de modelos de aprendizaje profundo. Un tensor core puede realizar operaciones matemáticas con matrices de números, que son estructuras que describen la relación entre otros objetos matemáticos. Estas operaciones son útiles para el cálculo de redes neuronales artificiales, que son el fundamento del aprendizaje profundo.

Los tensor cores pueden trabajar con diferentes precisiones, como FP32, TF32, FP16, INT8, INT4 y bfloat16, lo que les permite adaptarse a las necesidades de cada aplicación y optimizar el rendimiento y la eficiencia. Los tensor cores también se pueden usar para la computación científica de alta precisión, como la que se realiza en el HPC (High Performance Computing). [9]

#### 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/>