

1. Paral·lelisme:

La CPU (Processador) és molt ràpida processant tasques seqüencials (una darrere l'altra), mentre que la GPU destaca pel processament paral·lel. Explica per què el processament paral·lel és molt més eficient que el seqüencial per a l'entrenament de xarxes neuronals.

L'entrenament de xarxes neuronals es basa en fer un programa que ha de executar una tasca amb una sèrie de requisits i una sèrie d'accions possibles, i la manera en la que aprèn es fent la mateixa cosa un pic, i un altre, i un altre.

Quantes més iteracions puguis fer a la vegada, més opcions de trobar una millora té la xarxa neuronal, així que si enlloc de tenir 50 processos i fer-los un darrera l'altre en podem fer 25 cada vegada amb una GPU enlloc de una CPU, podem fer moltés més iteracions amb el mateix espai de temps.

2. Arquitectura Clau:

Quin element o nucli especialitzat que es troba a les GPU de NVIDIA és essencial per accelerar les operacions de càlcul de les matrius i tensors que es fan servir en Deep Learning?

Son els nucls Tensor, que estan present a partir de les gràfiques de serie 2000, i sobretot s'han emprat per poder donar pas a tecnologies com el DLSS (Deep Learning Super Sampling), una tecnologia que permet processar jocs a una resolució més baixa que la nativa per després tornar a pujar-la, per obtenir més fotogrames per segon
El meu portàtil, per exemple, té una 4060 i compta amb 96 Tensor Cores

GPU	
	NVIDIA GeForce RTX 4060 Laptop (AD107M/GN21-X)
	NVIDIA GeForce RTX 4060 Laptop
	AD107M/GN21-X4
	PCIe v4.0 x8 (16.0 GT/s) @ x8 (2.5 GT/s)
GPU #0	8.00 GB
	GDDR6 SDRAM
	128-bit
ROPs / TMUs	48 / 96
	SH/RT/TC
	3072 / 24 / 96

3. Llenguatge / Plataforma:

Quina plataforma de programació propietària de NVIDIA permet als desenvolupadors aprofitar directament la capacitat de càlcul paral·lel de les seves GPU per a tasques com el Deep Learning?

És la plataforma CUDA (Compute Unified Device Architecture)