GPU en Deep Learning

Patricio Cerda Mardini y Felipe Gómez Quezada

21 de noviembre de 2018

1 *GPU*

Nvidia "inventó" la GPU en el año 1999. Originalmente se concibió como un procesador de gráficos para PC. Era utilizado para el renderizado de imágenes, con énfasis en videojuegos y animación digital. Esa fue su única función conocida por muchos años.

2 Aparición de *GPU* en *Deep Learning*

Uno de los principales problemas que surgieron en el uso de *Deep Learning* como forma de solucionar problemas era la gran cantidad de cálculos que se requieren para poder entrenar un modelo. El tiempo de ejecución de los algoritmos era tan grande que las soluciones no eran factibles en la práctica.

En el año 2007, nVidia lanzó su *Compute Unified Device Architecture*, conocida por las siglas CUDA, una plataforma que, entre otras cosas, provee una librería que permite a los desarrolladores ocupar la capacidad de procesamiento en paralelo de la *GPU* para un propósito general. Ya no solo se ocuparía *GPU* para videojuegos o renderizado de modelos 3D, sino que también podría utilizarse para aplicaciones de cómputo intensivo. Gracias a esto se comenzó a utilizar *GPU* para el entrenamiento de redes neuronales profundas.

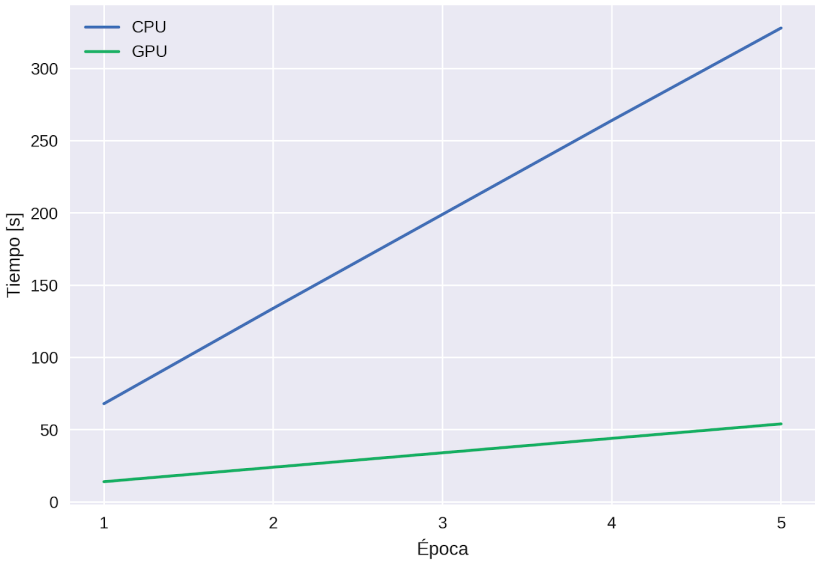
A pesar de que estaba la opción de usar *GPU*, nadie tenía la idea de que podía ser de utilidad, hasta que empezaron a aparecer investigadores que reportaron mejoras asombrosas en los tiempos de entrenamiento de redes neuronales profundas gracias al uso de procesamiento a través de *GPU*. Un claro ejemplo de esto fue el paper *Large-scale Deep Unsupervised Learning using Graphics Processors*, del año 2009, en donde se reportaron *speedups* de hasta 72 veces el tiempo sin uso de *GPU*. Redes que antes demoraban semanas en entrenarse ahora solo tardaban un par de horas. Esto significó una tremenda revolución en el valor de las *GPU's*, pues ahora no solo se concebía para jugar, sino como una poderosa fuente de poder de cómputo. Así nació el concepto de *GPGPU*, sigla del inglés General Purpose Computing on Graphic Processing Units (Cómputo

de propósito general en *GPU's*).

En 2014 Nvidia lanzó *cuDNN*, otra librería que sigue la línea de lo hecho en CUDA, esta vez favoreciendo implementaciones eficientes de las técnicas más populares utilizadas en *deep learning*, como *forward convolution*, *backward convolution*, *pooling*, *normalization*, entre otras.

3 Comparación en el rendimiento de *CPU* vs *GPU* en problema de clasificación de dígitos

En la demo mostrada se logra un *speedup* de 6x, lo que en una tarea tan sencilla de conseguir puede no ser muy relevante, pero en casos donde los modelos son más complejos, y los tiempos de entrenamiento del orden de semanas o meses, el speedup puede significar la diferencia entre un modelo útil y uno computacionalmente intratable.



4 Referencias bibliográficas

- Fecha de consulta: 21 de noviembre de 2018. Disponible en (URL): https://www.quora.com/Who-introduced-GPU-to-deep-learning
- *Large-scale Deep Unsupervised Learning using Graphics Processors*. Fecha de consulta: 21 de noviembre de 2018. Disponible en (URL): http://www.machinelearning.org/archive/icml2009/papers/218.pdf
- Fecha de consulta: 21 de noviembre de 2018. Disponible en (URL): https://developer.nvidia.com/cuda-gpus
- Fecha de consulta: 21 de noviembre de 2018. Disponible en (URL): https://arxiv.org/abs/1410.0759