GPGPUs in AI/ML

Overview

GPGPUs are very widely used in AI/ML, they can be said to be the standard computational tool for AI and ML problems, which lend themselves very well to GPGPU usage. GPUs are, predictably, used to make AI/ML programs run faster.

GPGPU usage in AI/ML

GPGPUs are not the only tool used for AI and machine learning, but they are the majority. NVIDIA calls GPUs "the de facto engines of AI computing" (they have a lot of good things to say about GPUs). A report by Epoch reported that "...most (if not all) of the biggest models over the last five years have been trained on GPUs or other special-purpose hardware…" Chat GPT is trained and run on GPUs. GPUs are used in both the training of models, and also something called "inference," which is when AI delivers the actual actionable insights. They're all over.

Why GPGPU?

GPU's are well suited to AI/ML tasks, they excel at doing just what makes up much of AI/ML, doing lots of similar linear algorithm calculations - especially true when talking about deep neural networks.

From Towards Data Science: "GPU for deep learning was a solution looking for a problem."

Applications

Training and inference of AI/ML models in any type of problem that they might be utilized for: image recognition, chat bots, etc.

Open Source Software

TensorFlow and PyTorch are examples of deep learning frameworks that support GPU acceleration. CUDA is used in AI/ML applications and so is cuDNN (CUDA Deep Neural Network) which is a library for deep neural networks.

Typically Implemented Algorithms

Convolutional Neural Networks

Recurrent Neural Networks

Generative Adversarial Networks

See many more here: Artificial Intelligence (AI) Algorithms - GeeksforGeeks

References:

A complete guide to AI accelerators for deep learning inference — GPUs, AWS Inferentia and Amazon Elastic Inference | by Shashank Prasanna | Towards Data Science

Why GPUs Are Great for AI | NVIDIA Blog

What Is AI Computing? | NVIDIA Blogs

What is Apache Spark? | Data Science | NVIDIA Glossary

Trends in GPU Price-Performance – Epoch (epochai.org)

ChatGPT (openai.com)

CUDA Deep Neural Network (cuDNN) | NVIDIA Developer