Title: Harnessing Computing Power for Machine Learning and Deep Learning: An In-Depth Review of Hardware, CUDA Core Architectures and their Impact on Performance

Author names: Md. Hasib Hasan; A.F.M Mahmud Ullah; Mohammed Anwer; Mahady Hasan; Md. Sadekur Rahman; Md. Tarek Habib

Abstract: Machine Learning (ML) and Deep Learning (DL) have emerged as transformative paradigms in modern computing, underpinning applications from natural language processing to autonomous systems and medical diagnostics. Central to their success is the computational acceleration provided by Graphics Processing Units (GPUs), particularly those based on NVIDIA's CUDA architecture. This study presents an in-depth analysis of CUDA-enabled GPU architectures spanning from Fermi to Hopper and their impact on ML/DL performance, scalability, and energy efficiency. Through detailed comparisons with traditional CPUs and specialized Tensor Processing Units (TPUs), we highlight the evolution of CUDA cores, memory hierarchies, and profiling tools that enable high-throughput, low-latency AI computation. Empirical studies, including convolution-heavy CNN tasks and real-time inference on edge devices, demonstrate substantial speedups and efficiency gains. We also explore GPU-specific optimizations such as kernel fusion, warp scheduling, and memory coalescing, emphasizing their role in accelerating training and inference. This study offers valuable insights into the hardware-software co-design strategies essential for scaling future AI workloads.

Keywords: Deep learning , Training , Tensors , Reviews , Scalability , Graphics processing units , Computer architecture , Hardware , Real-time systems , Optimization

DOI: 10.1109/ICMCSI67283.2026.11412792

https://ieeexplore.ieee.org/document/11412792/keywords#keywords