Revolutionizing machine learning: Harnessing hardware accelerators for enhanced AI efficiency

Cong Zou

College of Engineering, The Ohio State University, Columbus, 43210, USA

Zou.525@osu.edu

Abstract. In recent decades, the field of Artificial Intelligence (AI) has undergone a remarkable evolution, with machine learning emerging as a pivotal subdomain. This transformation has led to increasingly complex algorithms and soaring data volumes, necessitating robust computational resources. Conventional central processing units (CPUs) are struggling to meet the demanding requirements of modern AI applications. In response to this computational challenge, a new generation of hardware accelerators has been developed to enhance the processing and learning capabilities of machine learning systems. Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), and Application Specific Integrated Circuits (ASICs) are among the specialized accelerators that have emerged. These hardware accelerators have proven instrumental in significantly improving the efficiency of machine learning tasks. This paper provides a comprehensive exploration of these hardware accelerators, offering insights into their design, functionality, and applications. Moreover, it examines their role in empowering machine learning processes and discusses their potential impact on the future of AI. By addressing current trends and anticipated challenges, this paper contributes to a deeper understanding of the dynamic landscape of hardware acceleration in the context of machine learning research and development.

Keywords: Artificial Intelligence (AI), Machine Learning, Hardware Accelerators, Computational Efficiency.

1. Introduction

Over the past few decades, there has been a remarkable surge in the development of Artificial Intelligence (AI), with machine learning emerging as a pivotal subfield within AI. This transformation has placed machine learning at the forefront of research and innovation. As data volumes continue to skyrocket and algorithms become increasingly intricate, the need for robust computational resources has become paramount [1]. Conventional central processing units (CPUs) are struggling to keep pace with the demands of contemporary AI applications.

In response to this computational challenge, a new generation of hardware accelerators has been conceived and crafted to augment the processing and learning capabilities of machine learning systems. These hardware accelerators, including Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), and Application Specific Integrated Circuits (ASICs), have been tailor-made to cater to the unique requirements of machine learning workloads [2, 3]. Their specialization allows for a significant enhancement in the efficiency of machine learning tasks.

In this paper, it embarks on an exploration of these computational accelerators, delving into their inner workings and capabilities. We aim to provide readers with a comprehensive overview of these technological powerhouses, shedding light on how they empower machine learning processes. Additionally, we will cast our gaze toward the horizon, offering insights into the prospective trends and challenges that lie ahead in the realm of hardware acceleration for AI.

2. Theoretical Foundations and Working Principles

2.1. The definition of machine learning

Machine learning is a pivotal domain within artificial intelligence, emphasizing the capacity of computers to assimilate knowledge from extensive datasets, thereby refining their operational efficacy. Rather than operating on pre-defined algorithms, these systems are designed to learn from data, akin to deriving insights from patterns. The triad underpinning this discipline comprises data, features, and models. Data is the foundational input, with vast quantities indispensable for robust training. Each data point is characterized by its features, essentially its discernible attributes [4]. Subsequently, the model acts as a structured framework, capturing and representing the inherent patterns and relationships gleaned from the data, facilitating informed predictions and decisions.

There are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the algorithm picks up new information from labelled training data by translating input and output examples into output, for learning that is unsupervised. In order to meaningfully organize the data, algorithms immediately learn patterns from unlabelled data. The process of reinforcement learning entails an agent learning to operate in a way that maximizes some idea of cumulative reward while gradually changing its approach.

2.2. Types of hardware accelerators

Various types of hardware accelerators have come to the fore to pursue computational efficiency. These accelerators are engineered to handle specialized computational chores, offering a substantial boost in performance compared to the CPU.

The Graphics Processing Units originally engineered for rendering graphics in video games GPUs have found significant utility in the realm of machine learning. Their architecture allows for multiple threads to be executed in parallel, which is particularly beneficial for deep-learning tasks. While not as fast as other specialized hardware, CPUs are generally more versatile and can handle a wider range of tasks. Their architecture is optimized for serial processing but can be used for lighter machine-learning tasks that don’t require massive parallelism [5]. Tensor Processing Units (TPUs) were developed by Google specifically for the optimization of neural network computations. These units feature an architecture explicitly tailored for machine learning, focusing on high-throughput and low-precision arithmetic [6]. This unique design offers a compelling, energy-efficient alternative to more general-purpose GPUs in certain applications. On a similar note, Application-Specific Integrated Circuits (ASICs) present a hyper-focused approach to hardware acceleration. These custom-built components are optimized for highly specialized tasks, thus delivering exceptional performance within their specific domain of application. Field-Programmable Gate Arrays (FPGAs) offer another angle, providing configurable hardware landscapes that can be fine-tuned for particular tasks post-manufacture. This allows for a flexible yet optimized computing environment [7]. The selection among these various hardware options often hinges on a blend of factors including the complexity of the machine learning algorithms in play, the volume of data to be processed, and the ever-important consideration of energy efficiency.

2.3. Hardware Accelerators in Different Types of Learning

2.3.1. Supervised Learning

GPUs are the most choice for supervised learning, notably for algorithmic like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Boasting parallel computational architecture, these GPUs are best in matrix operations, data throughput, and transformational agility [8]. For example, NVIDIA's Tesla V100 has 640 Tensor Cores, each facilitating many simultaneous calculations.

2.3.2. Unsupervised Learning

Contrariwise, in unsupervised learning schema characterized by high-dimensional data manipulation, such as dimensionality reduction algorithms and clustering paradigms, Tensor Processing Units (TPUs) and Field-Programmable Gate Arrays (FPGAs) emerge as the hardware accelerators of preeminent feasibility. TPUs, endowed with custom-architected VLSI designs tailored for tensor computations, excel in hyperdimensional algorithms such as autoencoders and Generative Adversarial Networks (GANs) [9]. FPGAs, characterized by their reconfigurable hardware lattice, offer enhanced adaptability and have thus been ubiquitously deployed in real-time edge computing instances for algorithms like K-means clustering.

2.3.3. Reinforcement Learning

The spatiotemporal computational intricacies immanent in reinforcement learning paradigms necessitate hardware solutions that can concomitantly optimize real-time processing and multi-iterative stochastic computations. Herein, Application-Specific Integrated Circuits (ASICs) encapsulate an unparalleled advantage [10]. Designed for low-latency and high-throughput operationalities, ASICs are optimally conducive for real-time decisional algorithms such as Q-Learning and Monte Carlo Tree Search.

3. Application analysis and optimization of hardware accelerator and machine learning

3.1. Application of Hardware Accelerator

A study group from IHP Leibniz-Institut für innovative Mikroelektronik developed a hardware accelerator for solar particle event prediction with supervised machine learning. They perfectly showed how supervised machine learning techniques combine with a hardware accelerator to predict the in-flight upset rate. They choose supervised learning because it trains the model through known input and output data to ensure the accuracy and reliability of the model’s prediction [11].

In their design, a hardware accelerator is linked with an SRAM-based SEU monitor, and it can predict and monitor SEU rates in real-time. Such real-time prediction is important because it can provide an early warning for the upcoming radiation level of the systems potentially impacted by radiation during solar particle events. To achieve this goal, they first processed and transformed hourly SER data from historical solar events and aligned it with actual upsets from the SEU monitor. This ensures the machine learning model was trained on data congruent with accurate monitor outputs. Based on the data they observed, they trained two supervised machine-learning models, the RNN and the Linear Least Square model. Although they found out the RNN model performs slightly better than the Linear Least model, the second one uses fewer computational resources in limited hardware environments and is simpler. This is why they use the Linear Least Square model for the hardware accelerator.

Their proposed approach involves integrating an SRAM-based SEU monitor with results derived from an offline-trained machine learning model. This innovative design leverages on-chip SRAM as a real-time particle detector and employs two separate register files. One file records real-time hourly SEU data from the monitor, while the other stores parameter results obtained through offline machine learning. Furthermore, they have incorporated an accumulator for performing necessary calculations, and the inputs and functionality of this accumulator are determined through a straightforward control logic. To streamline the process and reduce hardware complexity, they have magnified coefficients within the function and retained only the integer part, simplifying the equation. Their design has been crafted with the intention of serving as an integral component for spaceborne systems. As such, simplicity and flexibility have been prioritized. In essence, their hardware accelerator design aims to deliver real-time SEU rate predictions, estimating the increase in radiation levels and mitigating the risk of exposing the target system to adverse conditions without adequate protection.

3.2. A case study in reinforcement learning optimization

The reinforcement learning is an important type of machine learning. Q-learning is one of the learning methods of reinforcement learning algorithms. The Q-learning employs a structure called Q-matrix, where each element reflects the predicted reward of performing a specific action in a specific condition. As the size of the problem grows, so do the computation and storage requirements of Q-learning. Spanò and his colleagues show how to combine Q-learning and hardware accelerator to make the algorithm more efficient [11].

Their proposed Q-learning agent comprises two main elements: the Policy Generator (PG) and the Q-learning accelerator. The PG is responsible for determining the next action by referencing the Q-matrix stored within the Q-Learning accelerator. This architectural choice enables the agent to make real-time decisions by considering both the current state and the Q-matrix values. In pursuit of creating a versatile accelerator, they intentionally omitted a predefined implementation for the PG, as its structure depends on the specific application. Nonetheless, they did incorporate the PG into their experiments to facilitate comparisons with established technologies.

The Q-Learning accelerator is the key point of their design. They store the Q-matrix in Z Dual-Port RAMs, labeled Action RAMs. Each RAM dedicated to an action encompasses an entire Q-matrix column. The number of memory slots mirrors the state count N. The system’s design makes sure that the read address corresponds to the next state st+1 while the current state st serves as the write address. A Q-matrix row, Q (st+1, A), is represented by the outputs of the Action RAMs, which are controlled by the current action.

Spanò and his team introduced several pivotal optimizations. First, approximated multipliers in Q-Updater. Even though the traditional multipliers are accurate, but they can be resource intensive. They proposed the use of approximated multipliers, leveraging barrel shifters. This not only makes the hardware simpler but also reduces power consumption. By approximating certain values to their nearest power of two, they managed to simplify calculations without compromising the Q-Learning algorithm’s convergence. Second, optimized MAX block. The Q-learning algorithm needs to identify the maximum Q-value for a particular state. Spanò and his team refined this process by deploying a tree structure of binary comparators, achieving a balance between speed and area efficiency. The third method is to make an efficient Q-matrix update equation. They restructured the Q-matrix’s update equation, making it possible to perform calculations using two multipliers instead of conventional threes. This strategic rearrangement further reduces hardware requirements and enhances the system's overall efficiency [12].

4. Future trends and challenges

4.1. Future trends

Integrating with quantum computation is one of the most notable directions regarding the future development of hardware accelerators. Integration of hardware accelerators with quantum computing can potentially solve complex machine learning problems, especially optimization tasks at unprecedented speeds. With the proliferation of IoT devices in the future, there’s a growing trend towards edge computing. Edge devices will increasingly incorporate hardware accelerators, allowing real-time machine-learning computations without communicating with a central server. As the demands for computation grow, the energy consumption will also increase. Therefore, energy-efficient designs would be an important task. Future hardware accelerators will prioritize energy efficiency and make sure that high-performance computations don’t come at the cost of excessive power usage. Another promising direction is neuromorphic computing [13]. By emulating the human brain's architecture, neuromorphic chips are designed to execute machine-learning tasks with heightened efficiency. These chips, processing information in a manner reminiscent of biological brains, could revolutionize the AI hardware domain.

4.2. Challenges

While there are a lot of future directions for hardware accelerators and machine learning, there are also a lot of good prospects. However, there are still a lot of challenges that should not be ignored. The first one is about Scalability. As machine learning models become increasingly complex, it is a challenge to ensure that hardware gas pedals can scale to meet the arithmetic requirements without losing performance. The cost of hardware gas pedals will also likely be a challenge. If AI devices are to be popularized in the future, it will be important to reduce costs to make cheaper products available to more small companies and designers. How to deal with heat dissipation will also be an issue. High-performance computing generates heat, affecting the device's performance and power consumption. Therefore, minimizing heat dissipation without affecting performance will be a challenge. The standardization could also be a challenge. It is important to reach standardization, which reduces costs and improves compatibility and operability of products on different platforms.

5. Conclusion

This paper delves into the theoretical foundations of machine learning and explores various hardware accelerators such as GPUs, TPUs, ASICs, and FPGAs, each customized for a specific learning paradigm. Through detailed case studies, this paper shows this accelerator's optimization strategies and applications in real-world scenarios, such as predicting solar particle events and augmenting Q-learning algorithms. The future directions of hardware accelerators, such as quantum computing integration, edge computing, energy-efficient design, and neuromorphic computing, are also clear. However, there are always challenges that come with development, such as scalability, cost, heat dissipation, and standardization, which are significant hurdles that the researchers in this area must overcome. In conclusion, hardware accelerators have greatly improved the efficiency of machine learning models, and future engineers and researchers require concerted efforts in research, design, and collaboration to address impending challenges and truly democratize AI for all.

References

[1] Lee J, Yoo H J. An overview of energy-efficient hardware accelerators for on-device deep-neural-network training. IEEE Open Journal of the Solid-State Circuits Society, 2021, 1: 115-128.

[2] Smagulova K, Fouda M E, Kurdahi F, et al. Resistive neural hardware accelerators. Proceedings of the IEEE, 2023, 111(5): 500-527.

[3] Mazzia V, Khaliq A, Salvetti F, et al. Real-time apple detection system using embedded systems with hardware accelerators: An edge AI application. IEEE Access, 2020, 8: 9102-9114.

[4] Bachem O, Lucic M, Krause A. Practical coreset constructions for machine learning. arXiv preprint arXiv:1703.06476, 2017.

[5] Ghimire D, Kil D, Kim S. A survey on efficient convolutional neural networks and hardware acceleration. Electronics, 2022, 11(6): 945.

[6] Xiao J, Andelfinger P, Eckhoff D, et al. A survey on agent-based simulation using hardware accelerators. ACM Computing Surveys (CSUR), 2019, 51(6): 1-35.

[7] Zhou X, Canady R, Bao S, et al. Cost-effective hardware accelerator recommendation for edge computing. 3rd USENIX Workshop on Hot Topics in Edge Computing (HotEdge 20). 2020.

[8] Zhao R, Luk W, Niu X, et al. Hardware acceleration for machine learning. 2017 IEEE computer society annual symposium on VLSI (ISVLSI). IEEE, 2017: 645-650.

[9] Du L, Du Y. Hardware accelerator design for machine learning. Machine Learning-Advanced Techniques and Emerging Applications. IntechOpen, 2017.

[10] Mittal S, Umesh S. A survey on hardware accelerators and optimization techniques for RNNs. Journal of Systems Architecture, 2021, 112: 101839.

[11] Chen J, Lange T, Andjelkovic M, et al. Hardware accelerator design with supervised machine learning for solar particle event prediction. 2020 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT). IEEE, 2020: 1-6.

[12] Spano S, Cardarilli G C, Di Nunzio L, et al. An efficient hardware implementation of reinforcement learning: The q-learning algorithm. Ieee Access, 2019, 7: 186340-186351.

[13] Wang C, Lou W, Gong L, et al. Reconfigurable hardware accelerators: Opportunities, trends, and challenges[J]. arXiv preprint arXiv:1712.04771, 2017.