Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 4s (2024)

Revolutionising Ai Deployment: Survey On Hardware Acceleration for Machine Learning Optimisation and Resilience

G. Pooja[1], Dr. S. Malathy[2]

[1] Department of AI&DS, Sri Shakthi Institute of Engineering and Technology, Coimbatore .

[2] Department of ECE, Karpagam academy of higher education Coimbatore.,

Article History:

Received: *21-04-2024*

Revised: *10-06-2024*

Accepted: *22-06-2024*

Abstract:

This compilation of research studies holds the utmost significance in hardware acceleration for machine learning. In our current era, characterised by the exponential growth of artificial intelligence (AI) applications, these studies tackle crucial challenges in optimising neural network accelerators' performance, energy efficiency, and resilience. The importance lies in their potential to revolutionise AI implementation across various domains. Efficient hardware accelerators are a cornerstone in unlocking the full potential of AI, enabling breakthroughs in deep learning, high-speed train fault detection and isolation, and numerous other applications. By improving memory management, facts placement, bus scheduling, and fault tolerance, that research paves the way for AI structures which are both powerful and sustainable, making AI accessible and impactful in a wide variety of fields. This research is important for fostering the growth and adoption of AI, ultimately remodelling how we interact with technology and facts in our daily lives.

Keywords: Artificial Intelligence, Accelerators, Machine Learning, Deep Learning, Energy Efficiency.

1. Introduction

Acceleration design, within the context of laptop engineering and hardware improvement, refers to enhancing the performance of specific computational duties or workloads. The primary objective of acceleration design is to achieve faster and more green execution of those responsibilities. This is often performed by utilising specialised hardware or accelerators designed to dump and expedite unique kinds of computation. Field-Programmable Gate Arrays (FPGAs) are a common hardware platform for acceleration layout. FPGAs are integrated circuits that provide unique flexibility and customisation for hardware acceleration. Unlike conventional processors (CPUs) and snapshot processors (GPUs), which have constant architectures, FPGAs can be programmed and reconfigured to carry out particular obligations [1]. This adaptability makes FPGAs surprisingly versatile for many packages, including acceleration in numerous domains and machine-learning knowledge of signal processing and cryptography [2].

FPGA and Accelerations layout

Field-Programmable Gate Arrays (FPGAs) are flexible hardware gadgets that play a crucial role in acceleration design. They consist of configurable logic blocks containing lookup tables, multiplexers, and flip-flops that can be programmed to implement specific logic features. These logic blocks are interconnected through a programmable interconnect network, permitting indicators to be routed between them. This interconnect material enables custom record paths and connections tailor-made

to the favoured computation. FPGAs are programmed by specifying the preferred logic capabilities and interconnections through a hardware description language (HDL) or a high-stage programming language. This programming configures the FPGA to behave as a devoted hardware accelerator for the favoured challenge[3].

FPGAs are prized for their potential to offer hardware acceleration custom-designed to specific workloads. They offer numerous benefits and the potential to create custom hardware circuits matching the computation required for unique tasks. This customisation can lead to great overall performance gains compared to running the same computations on well-known CPUs. FPGAs are also rather parallelisable, able to execute multiple operations simultaneously, making them appropriate for duties that may be divided into smaller, unbiased computations[4]. Additionally, FPGAs may be designed for strength performance, making them properly applicable for programs wherein power intake is an important challenge. Their reconfigurability and project-particular optimisation allow them to provide low-latency processing, vital in applications requiring real-time or close to real-time performance. Furthermore, their adaptability allows them to be reprogrammed to accommodate converting workloads or new algorithms, making them flexible answers for acceleration in dynamic domains of artificial intelligence and records processing. In summary, FPGAs are instrumental in acceleration design, imparting flexibility, customisation, parallelism, electricity efficiency, low latency, and adaptableness for unique computational responsibilities[5].

2. Highlights in these surveys

These surveyed work together encompass many subjects within hardware acceleration and neural networks, dropping mild at the ever-evolving landscape of device mastering hardware optimisation. They delve into the intricacies of designing green hardware accelerators tailored for edge processing, balancing high computational performance and minimum strength consumption. The demanding situations related to reminiscence get admission to in deep neural community applications are explored within the context of various domain names, starting from the Internet of Things (IoT) to cloud computing and records centres [6].

Custom accelerator design takes the middle stage as researchers emphasise the significance of correctly dealing with huge-scale Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs), specifically within the context of memory optimisation. Integrating neural accelerators with Graphics Processing Units (GPUs) is a first-rate pursuit, aiming to decorate basic performance and efficiency. Introducing Sparse CNN (SCNN) accelerators that make the most sparsity in neural networks gives a promising street for improved overall performance and electricity efficiency[7].

Neuromorphic photonic engines harness the potential of mild-based technologies and modern architectures to make a bigger computing overall performance. Accelerating approximate packages via neural networks emerges as a method to decorate overall performance and electricity efficiency[8] noticeably. Hardware Artificial Neural Network (ANN) accelerators for heterogeneous multi-middle systems are explored, highlighting their illness tolerance and strength performance and illustrating their ability in this context.

In an exceptional domain, detecting faults in high-velocity educate suspension systems is tackled

using a facts-pushed method grounded in a neural network[9]s. The position of Field-Programmable Gate Arrays (FPGAs) in accelerating ML knowledge of packages is very well tested, offering insights into their performance in contrast to traditional CPUs and GPUs.

Furthermore, the creation of noise-resilient deep Learning photonic neural community layouts running at high compute costs showcases their capacity to gain amazing accuracy, exceeding 99% on the MNIST dataset. Collectively, these works represent a large tapestry of research efforts committed to advancing the field of gadget-mastering hardware, reaffirming the dynamic and innovative nature of this unexpectedly evolving area[10].

3. Accelerations design and machine learning

Acceleration design within the context of system studying is a multifaceted field that involves improving specialised hardware and software components to enhance the overall performance and efficiency of machines getting to know responsibilities. ML knowledge, particularly deep learning, has seen a tremendous surge in programs throughout numerous domains, such as computer vision, natural language processing, speech popularity, self-reliant structures, and more. However, the computational needs of those gadget learning algorithms have also grown exponentially, requiring modern answers to satisfy performance and energy performance targets[11].

The number one goal of acceleration layout is to expedite the execution of ML algorithms[12], which frequently involve complicated mathematical operations like matrix multiplications, convolutions, and nonlinear activations. Several approaches and technologies are employed to attain this acceleration:

Hardware Accelerators: Specialised hardware components are designed to perform gadget mastering tasks more correctly than trendy-motive CPUs[13]. These hardware accelerators encompass Graphics Processing Units (GPUs), Field-Programmable Gate Arrays (FPGAs), and Application-Specific Integrated Circuits (ASICs). They offer high parallelism and optimised architectures for the specific requirements of system learning models. GPUs, for instance, have emerged as famous for deep mastering education due to their potential to carry out many parallel matrix operations simultaneously.

Custom Software Libraries: Optimised software program libraries and frameworks are created to maximise to-be-had hardware resources[14]. These libraries are specially tailor-made to leverage the competencies of CPUs, GPUs, or other accelerators. They permit ML algorithms to run quicker on fashionable hardware. Examples include the CUDA platform for NVIDIA GPUs and Intel's MKL for CPUs.

Approximation Techniques: In certain instances, gadget mastering models and computations may be approximated to reduce complexity. These approximation strategies aim to maintain suitable tiers of accuracy, extensively enhancing execution velocity and strength efficiency. Techniques like quantisation (lowering the precision of facts) and pruning (eliminating needless model parameters) fall under this category[15].

Quantisation and Pruning: Machine studying models can be quantised to lower bit precision, reducing reminiscence and computation necessities. Pruning entails figuring out and casting off less

applicable parameters in neural networks, making them more green without sacrificing overall performance.

Distributed Computing: Machine studying responsibilities may be accelerated by deploying them on dispensed computing structures with clusters or cloud resources. These systems enable parallel processing of facts and computations, decreasing education and inference times[16].

Hybrid Systems: Combining multiple hardware and software program additives, including CPUs, GPUs, and FPGAs, into hybrid structures allows for more green acceleration of ML duties. These systems harness the strengths of different accelerators to optimise overall performance at the same time as managing strength intake.

The significance of acceleration layout in device learning lies in its capability to meet the growing call for real-time processing, permitting side computing, and facilitating the coping with massive records and complicated fashions. Applications ranging from photo and speech reputation to recommendation systems, independent vehicles, and fraud detection have advantages from acceleration strategies. Researchers and engineers are continuously exploring progressive methods to improve acceleration layout, thereby advancing the competencies of machine learning throughout diverse domain names. The ultimate intention is to bridge the gap between the computational needs of the ML of fashions and the to-be-had hardware sources, making sure that ML algorithms may be executed efficiently and correctly.

4. Literature review

In ref [17], we look at delving into the realm of neural community accelerator optimisations. It examines various optimisations normally used in the latest research. The goal is to create a complete toolbox of optimisation techniques, followed by quantitative records concerning their effects. By providing this toolkit, it becomes less complicated for hardware designers to choose and put in force the maximum suitable optimisations for their unique neural network accelerator design. These optimisations can cause vast memory financial savings and electricity reductions.

Ref [18] introduces a unique approach for enhancing fact placement in deep learning accelerators. It addresses the challenges of reminiscence entry to problems while processing deep neural networks. Implementing a bus scheduling strategy for information placement on dispensed nearby buffers streamlines information-waft mapping, distribution, and transfer processes. The look demonstrates how this technique complements runtime and minimises bandwidth requirements, specifically in scenarios involving famous deep studying networks like AlexNet, VGG-sixteen, and GoogLeNet.

In ref[19], it specialises in the design of an accelerator specialised in handling big-scale Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs). A unique factor of this research is its precise emphasis on the role of reminiscence in accelerator design. It highlights how reminiscence optimisation can impact the overall design, performance, and power efficiency of hardware accelerators for neural networks. The studies demonstrate that it's viable to design an accelerator that can turn in high throughput while keeping low energy consumption, making it suitable for an extensive range of packages.

In ref[20], a look explores using low-precision (INT8) trendy matrix multiplication (GEMM) for

accelerating CNN inference on cellular devices and the usage of systolic arrays. The studies go beyond the traditional systolic array architecture by introducing the idea of Tensor-PEs, which will increase operand reuse and data path performance. The resulting Systolic Tensor Array (STA) microarchitecture suggests vast enhancements in vicinity and strength performance when processing INT8 operands. This optimisation is mainly treasured for CNN inference on mobile and area gadgets.

Ref [21] addresses the project of idle inference cycles in deep studying accelerators. It proposes a custom Equinox accelerator that incorporates each inference and schooling service. One of the important thing advantages is its capacity to hold provider-level latency constraints for inference while utilising idle cycles for education. The study showcases the massive increase in throughput compared to dedicated education accelerators while reaching near-optimal latency for inference.

T In ref[22], the consciousness of this painting is on designing an illness-tolerant hardware neural community accelerator. It is based on the expertise that artificial neural networks (ANNs) showcase intrinsic error tolerance. The studies demonstrate the interpretation of this error tolerance into realistic illness tolerance for hardware neural network accelerators. By growing spatially extended ANNs, the research suggests that these accelerators can maintain functionality even in defective transistors. This introduces the ability for electricity-green and fault-tolerant accelerators in heterogeneous multi-centre structures.

In ref[23], look addresses the vital venture of fault detection in high-pace teach suspension structures. It evaluates existing methods for fault detection, categorising them into model-based and statistics-pushed approaches. The research introduces a singular 1D convolution network-based totally fault diagnostic method, emphasising robustness through techniques just like the Gaussian white noise strategy and facet sample training. The proposed method is evaluated via multibody dynamics simulations, demonstrating its efficacy in real-global scenarios.

Ref [24] explores the capability of incorporating photonic neuromorphic hardware for deep gaining knowledge. The look recognises photonics' fantastic attributes regarding electricity efficiency and high bandwidth. It combines those benefits with deep mastering by designing a photonic neural network that operates at excessive compute fees. To ensure robustness, the research makes a speciality of each hardware architecture and facts-conscious deep gaining knowledge of schooling. The experimental effects verify the photonic hardware's high-velocity and excessive accuracy abilities.

Ref [25] investigates the performance of emerging FPGA frameworks for accelerating convolutional neural networks. It takes a practical approach by benchmarking those frameworks on numerous FPGA gadgets and comparing the effects to CPUs and GPUs. By comparing both FPGA overall performance and performance, the observation presents insights into the capability of FPGAs for accelerating CNNs in actual-world applications.

In ref[26], it introduces a coherent photonic neural community able to run at high compute costs. What makes this research wonderful is its focus on noise-resilient training models. The photonic neural network is designed as a silicon photonic chip and achieves extraordinary accuracy in tasks which include the MNIST category. The computation fees are substantially higher than conventional digital opposite numbers, providing insights into the future of photonic-based deep learning

hardware.

In ref[27], the Sparse CNN (SCNN) accelerator structure is offered as a progressive solution for improving deep learning applications' performance and strength performance. It specifically capitalises on the prevalence of 0-valued weights and activations in deep neural networks. By implementing a novel dataflow that keeps sparse representations, the SCNN achieves full-size upgrades in overall performance and electricity efficiency. This method contributes to the efficient deployment of deep mastering algorithms.

Ref [28] emphasises the want for energy-green hardware acceleration for machine learning. It explores the benefits of neuromorphic architectures and their ability to outperform traditional computing hardware. By supplying a neuromorphic structure tailor-made for gadget mastering obligations, the research paves the manner for hardware nicely suited to rising AI workloads.

Ref [29] makes a speciality of optimising statistics movement inside deep getting-to-know accelerators. It addresses the assignment of minimising information switch times and enhancing data placement. By introducing green statistics motion strategies that use dispensed local buffers, they look at streamlining the general data drift. The outcomes show that these optimisations can cause giant reductions in runtime and bandwidth requirements.

Ref [30] introduces a low-overhead neurally extended architecture designed for GPUs. One of its key functions is the careful attention to energy efficiency, ensuring that the implementation is optimised for low-energy operation. By leveraging disbursed neighbourhood buffers and batching optimisation, this method improves each performance and efficiency while preserving low overhead, making it suitable for more than a few programs.

In ref[31], observe provides Tanji, a deep learning accelerator offering a unified crossbar structure. What sets Tanji apart is its versatility in effectively handling each convolutional and fully connected neural network. It showcases how this crossbar structure maximises facts reuse and on-chip statistics movement whilst processing convolutional layers, after which it leverages high-speed off-chip information access for completely-linked layers. The prototype accelerator is carried out on the Xilinx platform, accomplishing real-time overall program performance, including item detection.

Ref [32] extends the traditional systolic array architecture by introducing the idea of Tensor-PEs and a unique block-sparse facts format known as density-certain block (DBB). The Systolic Tensor Array (STA) microarchitecture leverages those innovations to enhance each location. Table 1 discusses the performance metrics and limitations of various acceleration designs.

Table 1 The Performance Metrics And Limitations Various Acceleration Design

Ref	Algorithm	Output	Limitation
Ref[17]	Neural Network	Area [um2] 1309 7998 Power [uW] 577 4230	potential to improve the energy efficiency and performance of edge machine learning accelerators
Ref [18]	Cnn	traffic distribution for GoogleNet, with approximately 42.29% and 19% improvements compared to a TPU (Tensor Processing Unit).	Memory data placement may introduce complexity and hardware dependencies.

Ref [19]	Nn	area (56%) and power consumption (60%). The NFU (functional units)	the accelerator excels at large-scale layers. It may be less efficient for handling the full range of layer sizes and types commonly encountered in real-world machine-learning workloads.
Ref [20]	Neural Gpu Accelerator Architecture	degradation in the quality of results to 2.5% 2.8x average energy reduction:	NGPU may be unsuitable for applications where even minor quality degradation is unacceptable.
Ref [21]	Deep Learning Accelerator (DLA With YOLO	consumes 440mW at a clock rate of 165MHz	DLA prototype performs well in the described setup. The real-world scalability of larger and more complex deep neural networks or deployments has yet to be discussed.
Ref [22]	General Matrix Multiplication (GEMM) To Optimise Convolutional Neural Network (CNN)	.14x reduction in area and a 1.97x reduction in power consumption	The benefits may vary depending on the specific CNN models and datasets used for inference.
Ref [23]	DNN Inference Accelerators	6.67x higher throughput, 13% power and 4% area	applicability of Equinox to other types of workloads or services is not covered
Ref[24]	Multilayer Perceptron Neural Network (NN)	latency is 54.4 energy consumption is 39 nanojoule power ranging from 79 μW to 277 μW.	The generalisation of these findings to different neural network models, applications, or hardware platforms may take a lot of work.
Ref [25]	SCNN Accelerator, Specifically Optimised For Cnns	improves performance by a factor of 2.7x and energy efficiency by a factor of 2.3x	The architecture relies on the effectiveness of network pruning during training to create sparse weights and activations in CNNs.

5. Discussion

The survey presented here delves into diverse elements of hardware acceleration for device learning, with a focal point on optimising neural network accelerators. These works collectively provide insights into specialised hardware's layout, efficiency, and capacity applications for deep studying. The survey of accelerator optimisations serves as a valuable aid for hardware designers, supplying a complete toolbox of optimisation techniques. This toolbox, subsidised via quantitative records on their results, simplifies choosing and imposing optimisations tailor-made to particular neural community accelerator designs. The said memory financial savings and power reductions exhibit the potential for developing efficient and high-performance accelerators.

Additionally, the research on bus scheduling and facts placement techniques addresses a critical assignment in deep learning accelerators: reminiscence access issues. By optimising information-go with the flow mapping and distribution, these strategies show super upgrades in runtime and bandwidth utilisation. This is specifically large within the context of famous deep learning networks like AlexNet, VGG-16, and GoogLeNet.

The heterogeneous multi-middle machine studying accelerator design introduces a clean angle by way of highlighting the crucial function of reminiscence in accelerator layout and performance. Achieving high throughput with low power consumption is vital for deploying present-day gadgets and getting to know algorithms in various systems, and this work offers a promising solution.

Moreover, the studies on fault-tolerant DNN accelerators pave the way for hardware resilient to defects, increasing the reliability and lifespan of neural network accelerators. This disorder-tolerant method aligns with the growing desire for durable, electricity-efficient hardware in heterogeneous multi-center systems. The look at photonic neuromorphic hardware indicates the capability of integrated photonics in handing over excessive-velocity and high-accuracy deep learning engines. The recognition of hardware-aware deep learning training fashions showcases the adaptability of those architectures to actual global applications[33].

FPGA acceleration for CNNs is gaining prominence with the upward thrust of AI workloads. This research demonstrates that FPGA frameworks can compete with GPUs in overall performance and performance, especially in scenarios like photo reputation and class. The consequences inspire the exploration of FPGAs as accelerators for CNNs.The coherent photonic neural network gives a glimpse into the future of high-speed deep mastering hardware. Its noise-resilient design and high compute costs give a compelling case for adopting photonic-based total accelerators in advanced AI packages[34].

The Sparse CNN accelerator, on the other hand, underscores the importance of exploiting zero-valued weights and activations for progressed overall performance and electricity performance. This concept aligns with the broader trend of optimising neural community hardware for green deployment. Together, These studies contribute to the growing frame of information in hardware acceleration for system mastering, providing treasured insights and realistic solutions for accelerating deep knowledge of responsibilities, even optimising power performance and fault tolerance. As system getting to know continues to affect various domains, these hardware accelerators hold promise in enhancing overall performance and allowing the deployment of AI programs throughout numerous systems and structures[35-41].

This research work has investigated the problem of Automatic Fundus Image Quality Assessment and Lesion Detection Algorithm for Diabetic Retinopathy Screening. Advancements in optics, microfabrication, digital sensors, and image processing have led to increasingly smaller, more powerful and more portable imaging devices. Integrating these devices into wireless networks makes possible secure image transmission for teleophthalmology applications. In future, this work will be integrated with fundus cameras and low-cost teleophthalmology applications for diabetic retinopathy screening.

6.Conclusion

In conclusion, this compilation of studies reveals the dynamic panorama of hardware acceleration for device studying, underscoring its important role in unleashing the whole capability of artificial intelligence. This research mirrors the consistent evolution and refinement of specialised hardware answers to optimise neural community accelerators.

The supplied survey of accelerator optimisations equips hardware designers with a complete

toolkit to decorate the performance of deep learning accelerators. By quantifying the results of diverse optimisation techniques it streamlines the technique of choosing and implementing solutions tailored to the precise desires of neural community accelerators. The verified memory savings and energy reductions emphasise the opportunity of making excessive-overall performance accelerators even as maintaining energy intake in check.

Furthermore, the investigations into records placement, bus schedules, and fault-tolerant designs deal with essential challenges in deep learning hardware. They provide realistic strategies to tackle memory, get the right of entry to troubles, decorate runtime, and ensure hardware resilience in the face of defects.

The layout of heterogeneous multi-centre ML knowledge of accelerators takes a critical step towards making modern device learning algorithms handy to a wide range of systems. By focusing on the position of memory, this study demonstrates that it's viable to create accelerators with excessive throughput even as retaining power performance, thereby expanding the horizons of machine learning packages.

The take a look at photonic neuromorphic hardware exemplifies the synergy between integrated photonics and deep learning. The progressive architecture showcases the high-pace and excessive accuracy skills of photonic-based accelerators. It introduces hardware-aware deep-learning knowledge of education models that may adapt to real-international constraints.

They have FPGA-primarily based acceleration for convolutional neural networks, positions FPGAs as promising contenders for AI workloads. The demonstrated performance and efficiency profits, particularly in photo recognition and category, spotlight the ability of FPGAs as value-powerful options to GPUs.

Alternatively, the coherent photonic neural community gives a glimpse into the destiny of hardware-elevated deep learning knowledge. Its robust, noise-resilient layout, blended with high-speed compute costs, recommendations at a new era in deep mastering acceleration, emphasising the power and performance of photonics.

Finally, the Sparse CNN accelerator advocates exploiting zero-valued weights and activations to boost overall performance and strength efficiency. This aligns with the continued attempt to optimise neural community hardware for greater efficient deployments in various packages. This research contributes to the ever-increasing realm of hardware acceleration for ML knowledge. They provide treasured insights, methodologies, and technologies that preserve and beautify neural network accelerators' performance and reliability. As synthetic intelligence will become increasingly more ubiquitous throughout diverse domain names and structures, these improvements empower the broader adoption of AI programs, beginning doorways to new opportunities and possibilities. Destiny holds the promise of even extra green and effective hardware solutions with a view to pressure the next technology of shrewd structures.

REFERENCE

[1] Wu, Y. N., Emer, J. S., & Sze, V. (2019, November). Accelerate An architecture-level energy estimation methodology for accelerator designs. In 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD) (pp. 1-8). IEEE.

[2] Chen, Y., Xie, Y., Song, L., Chen, F., & Tang, T. (2020). A survey of accelerator architectures for deep neural networks. Engineering, 6(3), 264-274.

[3] Cohen, S. L., Bingham, C. B., & Hallen, B. L. (2019). The role of accelerator designs in mitigating bounded rationality in new ventures. Administrative Science Quarterly, 64(4), 810-854.

[4] Cohen, S., Fehder, D. C., Hochberg, Y. V., & Murray, F. (2019). The design of startup accelerators. Research Policy, 48(7), 1781-1797.

[5] Parashar, A., Raina, P., Shao, Y. S., Chen, Y. H., Ying, V. A., Mukkara, A., ... & Emer, J. (2019, March). Timeloop: A systematic approach to dnn accelerator evaluation. In 2019 IEEE International Symposium on performance analysis of systems and software (ISPASS) (pp. 304-315). IEEE.

[6] Yan, M., Deng, L., Hu, X., Liang, L., Feng, Y., Ye, X., ... & Xie, Y. (2020, February). Hygcn: A gcn accelerator with hybrid architecture. In 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA) (pp. 15-29). IEEE.

[7] Peng, X., Huang, S., Luo, Y., Sun, X., & Yu, S. (2019, December). DNN+ NeuroSim: An end-to-end benchmarking framework for compute-in-memory accelerators with versatile device technologies. In 2019 IEEE International Electron Devices Meeting (IEDM) (pp. 32-5). IEEE.

[8] Sapra, N. V., Yang, K. Y., Vercruysse, D., Leedle, K. J., Black, D. S., England, R. J., ... & Vučković, J. (2020). On-chip integrated laser-driven particle accelerator. Science, 367(6473), 79-83.

[9] Yakimenko, V., Alsberg, L., Bong, E., Bouchard, G., Clarke, C., Emma, C., ... &Yocky, G. (2019). FACET-II facility for advanced accelerator experimental tests. Physical Review Accelerators and Beams, 22(10), 101301.

[10] Deng, L., Li, G., Han, S., Shi, L., & Xie, Y. (2020). Model compression and hardware acceleration for neural networks: A comprehensive survey. Proceedings of the IEEE, 108(4), 485-532.

[11] Reuther, A., Michaleas, P., Jones, M., Gadepally, V., Samsi, S., & Kepner, J. (2019, September). Survey and benchmarking of machine learning accelerators. In 2019 IEEE high performance extreme computing conference (HPEC) (pp. 1-9). IEEE.

[12] Lu, L., Xie, J., Huang, R., Zhang, J., Lin, W., & Liang, Y. (2019, April). An efficient hardware accelerator for sparse convolutional neural networks on FPGAs. In 2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM) (pp. 17-25). IEEE.

[13] Liang, S., Wang, Y., Liu, C., He, L., Huawei, L. I., Xu, D., & Li, X. (2020). Engn: A high-throughput and energy-efficient accelerator for large graph neural networks. IEEE Transactions on Computers, 70(9), 1511-1525.

[14] Geng, T., Li, A., Shi, R., Wu, C., Wang, T., Li, Y., ... &Herbordt, M. C. (2020, October). AWB-GCN: A graph convolutional network accelerator with runtime workload rebalancing. In 2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO) (pp. 922-936). IEEE.

[15] Yang, Y., Huang, Q., Wu, B., Zhang, T., Ma, L., Gambardella, G., ... & Keutzer, K. (2019, February). Synergy: Algorithm-hardware co-design for convent accelerators on embedded FPGAs. In Proceedings of the 2019 ACM/SIGDA international symposium on field-programmable gate arrays (pp. 23-32).

[16] Xiao, T. P., Bennett, C. H., Feinberg, B., Agarwal, S., & Marinella, M. J. (2020). Analog architectures for neural network acceleration based on non-volatile memory. Applied Physics Reviews, 7(3).

[17] Jokic, P., Azarkhish, E., Bonetti, A., Pons, M., Emery, S., & Benini, L. (2022). A construction kit for efficient low power neural network accelerator designs. ACM Transactions on Embedded Computing Systems (TECS), 21(5), 1-36.

[18] Mirmahaleh, S. Y. H., Reshadi, M., Bagherzadeh, N., &Khademzadeh, A. (2021). Data scheduling and placement in deep learning accelerator. Cluster Computing, 24(4), 3651-3669.

[19] Chen, T., Du, Z., Sun, N., Wang, J., Wu, C., Chen, Y., & Temam, O. (2014). Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. ACM SIGARCH Computer Architecture News, 42(1), 269-284.

[20] Yazdanbakhsh, A., Park, J., Sharma, H., Lotfi-Kamran, P., & Esmaeilzadeh, H. (2015, December). Neural acceleration for GPU throughput processors. In Proceedings of the 48th international symposium on microarchitecture (pp. 482-493).

[21] Zhu, H., Wang, Y., & Shi, C. J. R. (2019). Tanji: A general-purpose neural network accelerator with unified crossbar architecture. IEEE Design & Test, 37(1), 56-63.

[22] Liu, Z. G., Whatmough, P. N., & Mattina, M. (2020). Systolic tensor array: An efficient structured-sparse GEMM accelerator for mobile CNN inference. IEEE Computer Architecture Letters, 19(1), 34-37.

[23] Drumond, M., Coulon, L., Pourhabibi, A., Yüzügüler, A. C., Falsafi, B., & Jaggi, M. (2021, October). Equinox: Training (for free) on a custom inference accelerator. In MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture (pp. 421-433).

[24] Roukhami, M., Lazarescu, M. T., Gregoretti, F., Lahbib, Y., & Mami, A. (2019). Very low power neural network FPGA accelerators for tag-less remote person identification using capacitive sensors. IEEE Access, 7, 102217-102231.

[25] Parashar, A., Rhu, M., Mukkara, A., Puglielli, A., Venkatesan, R., Khailany, B., ... & Dally, W. J. (2017). SCNN: An accelerator for compressed-sparse convolutional neural networks. ACM SIGARCH computer architecture news, 45(2), 27-40.

[26] Chen, Y. H., Krishna, T., Emer, J. S., & Sze, V. (2016). Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE journal of solid-state circuits, 52(1), 127-138.

[27] Esmaeilzadeh, H., Sampson, A., Ceze, L., & Burger, D. (2012, December). Neural acceleration for general-purpose approximate programs. In 2012 45th annual IEEE/ACM international symposium on microarchitecture (pp. 449-460). IEEE.

[28] Temam, O. (2012). A defect-tolerant accelerator for emerging high-performance applications. ACM SIGARCH Computer Architecture News, 40(3), 356-367.

[29] Chen, T., Chen, Y., Duranton, M., Guo, Q., Hashmi, A., Lipasti, M., ... & Temam, O. (2012, November). BenchNN: On the broad potential application scope of hardware neural network accelerators. In 2012 IEEE International Symposium on Workload Characterization (IISWC) (pp. 36-45). IEEE.

[30] Temam, O. (2012). A defect-tolerant accelerator for emerging high-performance applications. ACM SIGARCH Computer Architecture News, 40(3), 356-367.

[31] Ye, Y., Huang, P., & Zhang, Y. (2022). Deep learning-based fault diagnostic network of high-speed train secondary suspension systems for immunity to track irregularities and wheel wear. Railway Engineering Science, 1-21.

[32] Moralis-Pegios, M., Mourgias-Alexandris, G., Tsakyridis, A., Giamougiannis, G., Totovic, A., Dabos, G., ... &Pleros, N. (2022). Neuromorphic silicon photonics and hardware-aware deep learning for high-speed inference. Journal of Lightwave Technology, 40(10), 3243-3254.

[33] Kljucaric, L., & George, A. D. (2023). Deep Learning Inferencing with High-Performance Hardware Accelerators. ACM Transactions on Intelligent Systems and Technology, 14(4), 1-25.

[34] Mourgias-Alexandris, G., Moralis-Pegios, M., Tsakyridis, A., Simos, S., Dabos, G., Totovic, A., ... &Pleros, N. (2022). Noise-resilient and high-speed deep learning with coherent silicon photonics. Nature Communications, 13(1), 5572.

[35] Zhao, Z., Zhang, S., Xu, Z., Bellisario, K., Dai, N., Omrani, H., & Pijanowski, B. C. (2017). Automated bird acoustic event detection and robust species classification. Ecological Informatics, 39, 99–108. doi:10.1016/j.ecoinf.2017.04.003

[36] L. -L. Zhang, Y. Jiang, Y. -P. Sun, Y. Zhang and Z. Wang, "Improvements Based on ShuffleNetV2 Model for Bird Identification," in IEEE Access, vol. 11, pp. 101823-101832, 2023, doi: 10.1109/ACCESS.2023.3314676.

[37] Geetanjali S. Mahamunkar, Arvind W. Kiwelekar, and Laxman D. Netak,Deep Learning Model for Black Spot Classification,International Journal of Performability Engineering,18,2020, doi = {10.23940/ijpe.22.03.p8.222230}

[38] Shobhanam Krishna and Sumati Sidharth},HR Analytics: Employee Attrition Analysis using Random Forest,year = {2022},journal = {International Journal of PerformabilityEngineering,doi = {10.23940/ijpe.22.04.p5.275281}

[39] Rajan Prasad Tripathi, Sunil Kumar Khatri, and Darelle Van Greunen},Relative Examination of Breast Malignant Growth Analysis Utilizing Different Machine Learning Algorithms,year = {2022}, International Journal of Performability Engineering},doi = {10.23940/ijpe.22.06.p4.417425}

[40] Poonam Narang, Ajay Vikram Singh, and Himanshu Monga},Hybrid Metaheuristic Approach for Detection of Fake News on Social Media},International Journal of Performability Engineering},doi = {10.23940/ijpe.22.06.p6.434-443}

[41] Yerriswamy T and Gururaj Murtugudde},Signature-based Traffic Classification for DDoS Attack Detection and Analysis of Mitigation for DDoS Attacks using Programmable Commodity Switche, International Journal of Performability Engineering},doi = {10.23940/ijpe.22.07.p8.529536}