Analysis the Principle of AI Chips and the State-of-art Applications

Qi Zheng

Institute of Advanced Study, Shenzhen University, Shenzhen, China

2021280586@email.szu.edu.cn

Abstract. In recent years, the artificial intelligence (AI) chips have emerged as indispensable components specialized for executing AI computations, which is considered as a new era for integrate circuit industry. On this basis, this study comprehensively examines the underlying principles, performance characteristics, diverse applications, limitations, and prospective directions in AI chip research. The parallel architecture, sparsity, quantization, and adaptability of AI chips empower the expeditious and efficient processing of AI tasks. These chips find versatile applications in computer vision, healthcare, and intelligent robotics, amplifying their transformative impact. Nevertheless, challenges persist in terms of scalability, versatility, and the rapid pace of technological evolution. As a matter of fact, ongoing research endeavors should prioritize the symbiotic optimization of hardware and software, the creation of energy-efficient designs tailored to specific domains, and the seamless integration of emerging technologies. These results offer a lucid depiction of the prevailing state and the anticipated trajectory of AI chip development.

Keywords: AI chips, neural networks, parallel architecture.

1. Introduction

The foundations of modern AI can be traced back to computers and mathematical theory in the early 20th century. British mathematician Alan Turing published a paper in 1950 that proposed the concept of the "Turing test": requiring a computer to behave like a human in a conversation with a human so that other humans cannot distinguish between a computer and a human interlocutor is a way to assess whether a machine is intelligent [1]. The history of chip development dates back to the mid-20th century, when early computers used vacuum tubes to build electronic circuits. The concept of an Integrated Circuit (IC) was introduced by Jack Kilby and Robert Noyce in 1958. Subsequently, in the 1960s to 1970s, with the progress of technology, the scale of integrated circuits continued to increase, from small scale integrated circuit (SSI) to medium scale integrated circuit (MSI), and then to large-scale integrated circuit (LSI) and very large-scale integrated circuit (VLSI), the number of transistors on the chip increased significantly, and the performance was greatly improved. In the 1990s and 2000s, SOCs developed and were widely used in mobile devices, embedded systems, and the Internet of Things. Today, the rise of multi-core processors and AI chips. With the development of artificial intelligence, AI chips have become a hot field. AI chips optimized specifically for AI computing tasks play an important role in deep learning and AI applications.

AI in many areas of life can play a role in advancing the industry. Meanwhile, AI Chips, a microprocessor designed specifically for AI tasks, could drive AI adoption in these industries. In the field of healthcare, AI chips tailored for cancer models offer notable advantages in the realm of preclinical drug screening. These chips are meticulously designed to enable efficient and high-throughput evaluation of various biological parameters, including antitumor drug response. This innovative approach facilitates rapid analysis of drug efficacy and its impact on tumor biology. The tailored architecture of these AI chips empowers researchers with a sophisticated tool to expedite preclinical assessments, contributing to the acceleration of drug development and personalized treatment strategies for cancer [2]. In the financial sector, as per the Business Insider Report published in June 2019, AI techniques are employed across three key sectors in banking: conversational banking, anti-fraud detection, and risk assessment, including credit underwriting. Moreover, AI-driven financial chatbots and voice assistants replicate human interactions, enhancing customer engagement

by providing tailored insights and suggestions. Furthermore, AI's widespread use in intelligent trading systems supports stock market and currency price forecasting, aiding decisions to buy, hold, or sell based on evolving indicators. Additionally, machine learning-driven anti-fraud mechanisms learn from transaction patterns to identify anomalies, reinforcing security measures [3]. The rapid inference capabilities of AI chips enable swift determination of transaction security and identification of potential fraud risks. The continuous learning feature of AI chips allows for ongoing optimization of trading algorithms based on new market data and trading strategies, thereby enhancing the performance and security of the trading system. In the field of transportation and intelligent transportation, artificial intelligence based on neural networks obtains transport data from vehicles by detecting different physical profile patterns for each mode, including speed, acceleration, number of satellites in the field of view, and electromagnetic levels [4]. AI chips can provide high-performance and powerful parallel computing capabilities, which are suitable for large-scale data processing and parallel computing of machine learning algorithms. This allows the analysis and prediction of traffic data to be completed in a relatively short period of time, providing timely feedback on traffic conditions. In this article, the author will analyze the principles and latest applications of AI chips.

2. Basic Descriptions of AI chip

The proliferation of artificial intelligence (AI) applications has underscored the need for hardware tailored to AI's unique computational demands. At the forefront of this hardware evolution are AI chips, specialized components engineered to accelerate and optimize AI tasks. AI chips, also known as AI accelerators, are meticulously designed to efficiently handle the computations inherent to AI and machine learning (ML) workloads. The chip schematic depicted in Fig. 1 is sourced from Nervana's AI chip [5]. Fig. 2 illustrates Intel's Loihi22 and CEA-Leti's DynapSEL chips, which possess self-learning capabilities based on synaptic plasticity. These diagrams offer a general representation of the architecture commonly found in AI chips. The cornerstone of AI chip design resides in its ability to execute complex mathematical operations and data manipulations that underpin AI algorithms. The architecture of an AI chip is an intricate amalgamation of components, each strategically fashioned to cater to distinct computational requirements.

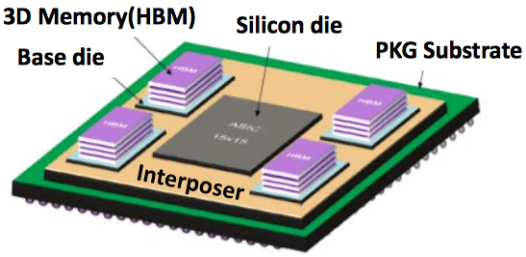


Figure 1. Nervana AI chip

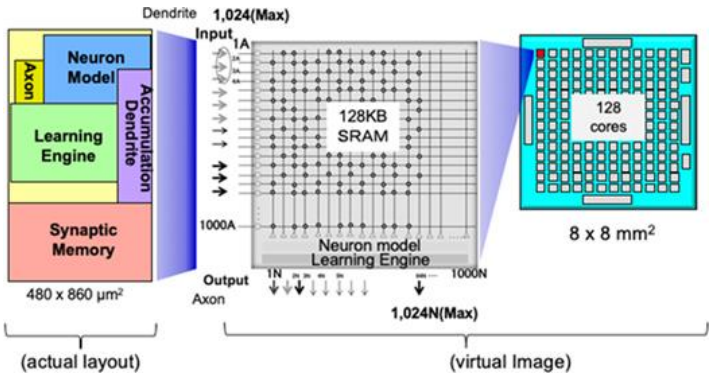


Figure 2. Loihi circuit diagram

Central to this architecture are processing units, the heart of AI chips. These encompass a spectrum of cores, ranging from versatile central processing units (CPUs) to parallel-processing graphical processing units (GPUs) and specialized application-specific integrated circuits (ASICs). CPUs offer adaptability across tasks, GPUs excel in parallel operations pivotal for neural network training, and ASICs deliver tailored performance enhancements for specific AI computations.

AI chip architecture extends beyond processing units to encompass memory hierarchies. These hierarchies incorporate high-speed caches and memory banks, ensuring efficient data access and minimizing latency during computations. The orchestrated interplay of data movement pathways and interfaces facilitates seamless communication between memory and processing units, mitigating potential bottlenecks and amplifying computational efficiency. Parallel architecture, a hallmark of AI chips, integrates multiple processing cores. This architecture permits concurrent execution of diverse tasks, addressing computational demands and expediting overall performance. Moreover, specialized arithmetic units embedded within AI chips optimize complex mathematical operations intrinsic to AI computations. Control logic units orchestrate the choreography of instructions and data flow within the chip, ensuring synchronized operations. The pursuit of energy efficiency stands as a pivotal aspect of AI chip design, particularly in the context of burgeoning edge computing and mobile applications [6]. In conclusion, AI chips epitomize the convergence of specialized hardware and AI algorithms. Their architecture, spanning processing units, memory hierarchies, and orchestrated data movement, encapsulates the essence of AI computation. This exploration provides an introductory glimpse into the intricate world of AI chips and their transformative role in shaping AI's trajectory.

3. Principle

In the realm of artificial intelligence (AI), the design and development of specialized hardware, known as AI chips, have emerged as pivotal pursuits. These chips are meticulously crafted to accommodate the computational demands inherent to AI and machine learning (ML) tasks. This exposé delineates the fundamental principles underlying the distinctive attributes of AI chips, focusing on parallel architecture, quantization, and neuron simulation. The core principle of combining parallel architecture and dedicated accelerators influences the design of AI chips. This approach enhances the efficiency of AI task execution, especially those involving intricate matrix operations. AI chips feature multiple processing cores for concurrent task handling, optimizing computational speed. These cores proficiently manage diverse operations simultaneously, substantially elevating computational efficiency. Additionally, hardware accelerators, like specialized matrix units, strengthen pivotal AI algorithms. Furthermore, AI chip architectures are tailored for neural algorithms' parallelism and distribution. They adopt a massively parallel computation paradigm, prioritizing inter-unit communication. Specific chips like TrueNorth [7] are highly specialized for a single neuron model with limited programmability. Others like SpiN-Naker [8] offer higher flexibility [9]. The amalgamation of parallel architecture and accelerators leverages parallelism's potential, facilitating swift intricate AI computations.

The pivotal principle of quantization serves as a cornerstone of AI-chip design, strategically embracing optimization techniques. Quantization operates by representing intricate mathematical operations, encompassing weights and activations within AI models, with lower bit-widths. This compression strategy curtails both storage and computation overhead. Leveraging the inherent resilience of neural networks to quantization noise, AI chips are meticulously crafted to execute operations using quantized parameters.

As an instance, the concept of weight sharing is depicted in Fig. 3, taking a layer comprising 4 input and 4 output neurons, yielding a 4×4 weight matrix. The upper-left quadrant showcases the weight matrix, while the lower-left quadrant exhibits the gradient matrix. Weight quantization into 4 bins (distinguished by colors) prompts weights within the same bin to share a value. Consequently, a compact index suffices to reference a shared weight value, curbing storage requirements. During updates, gradients are grouped by color and collectively summed. Though 16 weights initially existed,

only 4 shared weights emerge—similar weights conglomerate, minimizing redundancy. The original storage demand of 16 weights, each occupying 32 bits, dwindles to 4 effective weights as shown in Fig. 3 (blue, green, red, and orange), each spanning 32 bits, coupled with 16 2-bit indices. This yields a compression ratio of 16 * 32 / (4 * 32 + 2 * 16) = 3.2 [10]. This practice fosters a dual effect: it mitigates memory demands and accelerates computations—crucial considerations in power-constrained scenarios.

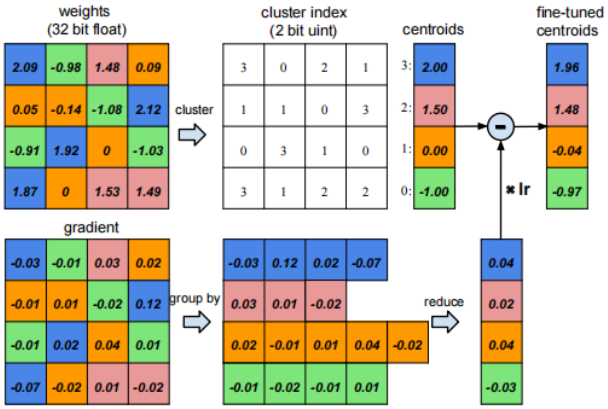


Figure 3. Weight sharing by scalar quantization (top) and centroids fine-tuning (bottom)

Neuron simulation, a foundational principle in AI-chip design, empowers emulation of human brain-like behavior. Accomplished through the integration of spiking neural networks, it mirrors biological neural impulses with discrete spikes that encode information. AI chips adeptly process temporal correlations, embracing neuron simulation to capture the dynamic attributes embedded within time-varying data. This principle's significance is particularly pronounced in edge AI domains, facilitating real-time processing of sensory inputs. The incorporation of neuron simulation underscores AI chips' capacity to simulate intricate cognitive functions for a spectrum of applications. In summation, AI-chip design principles are multifaceted, encompassing parallel architecture, quantization, and neuron simulation, alongside various other cutting-edge technologies. These principles engender the distinctive attributes of AI chips, endowing them with the capacity to execute intricate computations with efficiency and finesse. The amalgamation of these principles engenders a potent paradigm shift, propelling AI-chip capabilities beyond conventional computing architectures. As the AI landscape continues to evolve, these principles remain pivotal in steering AI-chip innovation towards heightened performance and efficiency.

4. Performance

AI chips have emerged as pivotal components in driving the advancement of artificial intelligence applications, showcasing distinctive performance characteristics that have garnered extensive attention within the academic community. These characteristics not only shape the efficiency and effectiveness of AI computations but also underscore their unique roles in specialized domains.

One standout feature lies in the emulation of Spiking Neural Networks (SNNs). AI chips have gained prominence for their ability to simulate SNNs, replicating the intricate pulse-based communication observed in biological neurons. This simulation approach makes them particularly suited for tasks requiring temporal correlation, such as pattern recognition and real-time edge computing. The emulation of SNNs represents a research-rich avenue within academia, delving into their role in understanding neural behavior, processing dynamic sensory data, and developing energy-efficient edge AI applications. IBM's well-known AI chip, TrueNorth [11], is adept at emulating Spiking Neural Networks (SNNs), which are characterized by their ability to facilitate spike transmission between neurons for event-driven and energy-efficient computational processes. The utilization of SNNs on neuromorphic hardware platforms offers the advantage of reduced energy consumption, making them a promising avenue for energy-efficient neural processing in various

applications. The TrueNorth architecture derives its inspiration from the intricate structure and functional dynamics of the human brain as shown in Fig. 4 (a), including the cortical column (b)—an essential unit composed of densely interconnected and functionally related neurons. This framework serves as the cornerstone for understanding brain functionality. Correspondingly, the neurosynaptic core (c), a fundamental module within the TrueNorth architecture, emulates this concept. It comprises 256 input axons, 256 neurons, and a 64k synaptic crossbar, mirroring the interplay of neurons, synapses, and communication pathways (c). This core integrates tightly coupled components: computation (neurons), memory (synapses), and communication (axons and dendrites). A singular core occupies a silicon area of 240μm × 390μm (d), encapsulating these functional elements. Scaling up, the assembly of 4096 neurosynaptic cores within a 2D array shapes the TrueNorth chip (e). This chip spans an area of 4.3cm² (f), created through a 28nm low-power CMOS process, and remarkably consumes only 65mW of power during the execution of a conventional computer vision task. The outcome is an effective emulation of cortical structure, skillfully executed within the confines of an inorganic silicon substrate (g). This architecture, by merging its modular components in line with brain-inspired principles, offers a promising avenue for advancing artificial intelligence and neuromorphic computing paradigms [11].

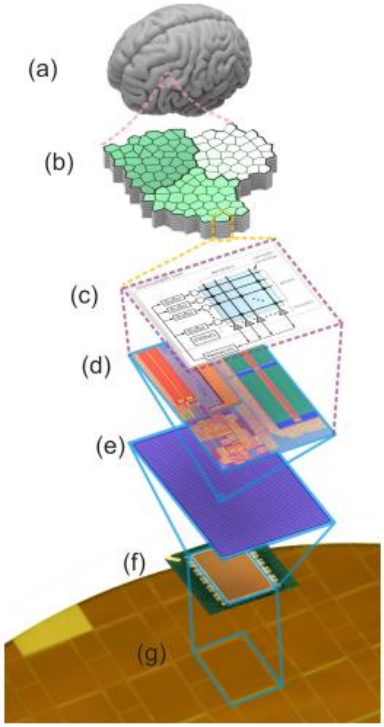


Figure 4. The TrueNorth architecture

The integration of sparsity and quantization forms another distinct characteristic of modern AI chips. These chips adopt methodologies that reduce computational load and memory demands while preserving model efficacy. In pursuit of heightened computational efficiency, an additional approach involves constraining the impact of weights on output. This structured form of sparsity emerges when each output is exclusively reliant on inputs from a predefined window. Furthermore, superior efficiency can be achieved through consistent application of identical weights throughout all output computations [12]. Termed "weight sharing," this concept markedly reduces the storage requisites for weight values, further refining resource allocation.

Simultaneously, the academic discourse has focused on investigating the effects of sparsity and quantization on neural network accuracy, energy efficiency, and effective hardware implementation. Researchers are driven by the imperative to strike a harmonious balance between computational intensity and resource consumption. This balance is explored intricately as they delve into the complexities of optimized model representations that can be effectively deployed across a spectrum of devices and scenarios. This holistic incorporation of techniques not only underpins the design of

contemporary AI chips but also empowers them to yield significant performance advantages, effectively harmonizing computational efficiency with model performance.

Furthermore, adaptability and reinforcement learning have emerged as distinctive facets of AI chips, resonating with the principles of continuous improvement. Some chips incorporate adaptability through reinforcement learning, allowing them to autonomously enhance their performance by learning from the environment and refining algorithms. The academic exploration of this area probes into the potential of these chips in dynamic and evolving contexts, encompassing adaptive decision-making, personalized assistance, and autonomous systems. This characteristic aligns with the broader drive to imbue AI systems with learning mechanisms that enable them to thrive in ever-changing environments.

The memristor-based modeling of the threshold characteristics of artificial nociceptors, as demonstrated in the study, highlights the significance of enhancing adaptability for artificial intelligence systems. The threshold sensitivity enables timely detection of harmful external stimuli exceeding preset limits, allowing the system to promptly respond and adapt to environmental changes. The post-relaxation hypersensitivity also maintains acute detection of new stimuli subsequent to responses. By emulating such adaptive capabilities of biological nociceptors, the programmable memristor-based implementation provides an approach to equip artificial intelligence systems with an adaptable sensory faculty. The bio-inspired feedback mechanism enhances perception, detection, and regulated response to varying external conditions. This promises to augment the adaptive aptitude and self-regulation abilities of intelligent systems to operate in unknown, dynamic environments. In conclusion, replicating biological adaptability mechanisms is instrumental for engineering resilient and context-aware artificial intelligence [13].

In conclusion, AI chips encapsulate a confluence of distinctive traits that define their contribution to the evolution of artificial intelligence. The emulation of Spiking Neural Networks, support for sparsity and quantization, and incorporation of adaptability through reinforcement learning stand out as hallmark features. These features stimulate deep academic inquiry, spanning domains such as neurocomputing, model optimization, energy efficiency, and adaptability in intelligent systems. As the field continues to evolve, these performance characteristics will undoubtedly play a pivotal role in shaping the future trajectory of AI research and applications.

5. Applications

AI chips have surfaced as integral and transformative elements that have found diverse applications across multiple domains. Notably, they have made significant contributions in fields such as Computer Vision, Healthcare, and Intelligent Robotics and Automation.

5.1. Computer Vision

In the realm of Computer Vision, AI chips leverage their high computational prowess to facilitate intricate tasks such as image recognition, object detection, and facial identification. Their accelerated image processing capabilities enable real-time visual analysis, exemplified by their deployment in autonomous vehicles for environmental perception and road identification, and in surveillance systems for event detection and tracking.

Taking the example of MoS2 to illustrate the applications of AI chips in computer vision, the study demonstrates the efficacy of MoS2 optoelectronic synapses in emulating biological visual processing and recognition functionalities. The large-scale integratability of MoS2 transistors enabled the fabrication of synaptic arrays to mimic neuromorphic vision sensing and preprocessing on AI chips. The light-tunability of MoS2 devices allowed them to exhibit essential synaptic characteristics, including paired-pulse facilitation and adjustable memory decay dynamics. By modulating the gate voltages, the MoS2 optoelectronic synapses successfully simulated interest-based tuning of human visual memory. Moreover, the synaptic arrays facilitated convolutional neural network-based image recognition with reduced noise and 85.5% accuracy on AI chips. Such programmable, photo

responsive emulation of neural mechanisms for computer vision can overcome constraints like complexity and redundancy in conventional digital systems on AI chips. The simple fabrication and synaptic versatility of MoS2 nanostructures underscore their potential as computational primitives for energy-efficient, bio-inspired artificial vision technologies on AI chips. In summary, this research highlights the unique benefits of two-dimensional MoS2 materials to advance beyond the limitations of prevailing computer vision paradigms by demonstrating their applications in AI chips [14].

5.2. Healthcare

Turning to the Healthcare domain, AI chips wield substantial potential for medical image analysis, disease diagnosis, and pharmaceutical research. These chips expedite the processing of extensive medical datasets, fostering accurate diagnosis and treatment planning by aiding healthcare professionals. Furthermore, they play a pivotal role in drug screening and genomic research, facilitating the identification of potential therapies and advancing precision medicine.

Microfluidics and AI are revolutionizing pharmaceutical research by rapidly fabricating precise drug delivery systems. AI chips play a pivotal role in this transformation. Microfluidic chips offer controlled microparticle fabrication, with AI predicting outcomes like particle size, reducing trial and error. The synergy between AI and microfluidics advances drug design and screening, enabling patient-specific medicine. Challenges in standardization and validation persist, but AI's integration with microfluidics accelerates drug discovery, toxicology, and precision medicine. Striving for clinical acceptance, AI-driven microfluidic platforms hold potential for cost-effective, personalized drug development [15].

5.3. Intelligent Robotics and Automation

In the realm of Intelligent Robotics and Automation, AI chips revolutionize the capabilities of robots and automated systems by enabling them to perceive their surroundings, make informed decisions, and execute tasks. Their integration in industrial automation scenarios empowers collaborative robots on production lines, optimizing manufacturing processes. Additionally, AI-powered service robots benefit from these chips, enhancing navigation, human-machine interaction, and cognitive capabilities, making them invaluable for various sectors.

Reinforcement learning and dedicated AI chips are driving advancements in intelligent autonomous systems and robotics. These AI chips enhance crucial computations like path planning, leading to rapid, adaptive decision-making for robots in complex environments, as the evaluation flow shown in Fig. 5. The integration of reinforcement learning enables robots to dynamically update navigation policies based on past experiences, enhancing obstacle avoidance. Specialized AI processors, like the one featured here, significantly accelerate key computations such as path planning, vastly outperforming general-purpose CPUs. This synthesis of AI techniques and specialized hardware facilitates agile, adaptive decision-making for autonomous robots in intricate real-world scenarios. Such integration holds significant potential for applications including surveillance, package delivery, and exploration [16].

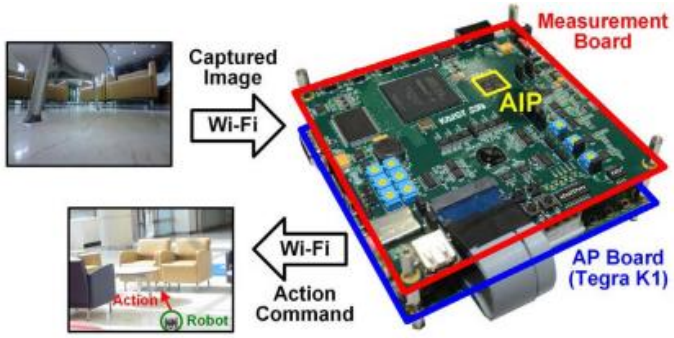


Figure 5. The diagram illustrates the Chip Evaluation System.

In summary, AI chips have ushered in a new era of technological innovation, transforming industries by enabling advanced capabilities in Computer Vision, Healthcare, and Intelligent Robotics and Automation. Their computational prowess, real-time processing, and efficient energy usage have collectively propelled the realms of image analysis, medical diagnosis, robotic intelligence, and more, heralding a future where AI-driven applications redefine the boundaries of what is possible.

6. Limitations & Prospects

The current landscape of AI-chip research presents certain limitations that warrant careful consideration. Firstly, the rapid pace of AI development has led to a disconnect between hardware and software optimization. While AI algorithms continue to evolve, the hardware architectures that support them may lag behind, inhibiting their full potential. Furthermore, the complexity of AI models poses a challenge in terms of scalability and efficiency on AI chips. Many contemporary models are large and resource-intensive, requiring substantial memory and computational resources that may exceed the capabilities of current AI-chip designs. Secondly, the trade-off between accuracy and energy efficiency remains a persistent constraint. As AI chips are often deployed in resource-constrained environments, achieving a balance between model accuracy and power consumption becomes crucial. Striking this balance requires innovative approaches in chip design, such as exploring new ways to optimize neural network architectures for better energy utilization without compromising performance. Another limitation lies in the domain-specific nature of AI chips. While AI chips excel in specialized tasks, they may lack the versatility to handle diverse AI applications. This restricts their adaptability and limits their potential impact across various domains. Furthermore, the field of AI is dynamic, and rapid advancements could render current AI-chip designs obsolete in a short span, emphasizing the need for flexible and upgradable architectures.

Despite these limitations, the future of AI-chip research is promising, offering several avenues for exploration and innovation. Hardware-software co-design is expected to play a pivotal role in addressing the existing challenges. Collaborative efforts between hardware engineers and AI researchers can lead to optimized architectures that exploit the unique characteristics of AI algorithms while overcoming hardware constraints. Moreover, there is a growing interest in domain-specific AI-chip designs that cater to the distinct requirements of various applications, enhancing both efficiency and performance. Another exciting prospect lies in neuromorphic computing, inspired by the human brain's architecture. Neuromorphic AI chips simulate neural networks with remarkable energy efficiency and parallel processing capabilities. Continued research in this field could lead to breakthroughs in energy-efficient AI processing, enabling the development of intelligent systems with reduced power consumption. Furthermore, advancements in materials science and fabrication technologies hold promise for creating more efficient and reliable AI chips. Novel materials and fabrication techniques could enhance chip performance, durability, and thermal management, addressing some of the current limitations. In the long term, the convergence of AI, quantum computing, and other emerging technologies could reshape the landscape of AI-chip research. Quantum AI chips, for instance, could potentially accelerate complex computations and enable breakthroughs in AI algorithms. In conclusion, while current AI-chip research faces challenges related to hardware-software optimization, energy efficiency, and domain-specificity, the field is poised for transformative growth. Collaborative efforts, innovative chip designs, and emerging technologies are expected to drive AI-chip research toward addressing these limitations and unlocking new opportunities for AI applications across diverse domains.

7. Conclusion

To sum up, this study delves deeply into the analysis of AI chip principles, performance, applications, and future research prospects. The underpinning principles of parallel architecture, quantization computations, and neuron simulation drive the efficiency and specialized capabilities of

AI chips. Notable features such as SNN emulation, sparsity, quantization, and adaptability empower AI chips to excel in high-performance AI tasks, though the ongoing optimization challenge involves balancing accuracy and efficiency. The transformative impact of AI chips is evident in domains like computer vision, healthcare, and intelligent robotics, enabling accelerated sensory processing, medical imaging, and robotic cognition. However, challenges related to hardware-software co-design, scalability, versatility, and rapid obsolescence persist.

Future research in AI chip development should prioritize collaborative hardware-software optimization, energy-efficient specialized designs, neuromorphic computing, and integration of emerging technologies. Such initiatives could enhance efficiency, scalability, and multifunctionality, unlocking the full potential of AI chips. In conclusion, this comprehensive review provides invaluable insights into the current state and future trajectory of AI chip research, highlighting their pivotal role in advancing AI capabilities while acknowledging inherent limitations as opportunities for innovation. The knowledge presented serves as a cornerstone for researchers in academia and industry pursuing advancements in this swiftly evolving field.

References

[1] Machinery C. Computing machinery and intelligence-AM Turing. Mind, 1950, 59 (236): 433.

[2] Fetah K L, DiPardo B J, Kongadzem E M, et al. Cancer modeling-on-a-chip with future artificial intelligence integration. Small, 2019, 15 (50): 1901985.

[3] Sabry F, Labda W, Erbad A, et al. Cryptocurrencies and artificial intelligence: Challenges and opportunities. IEEE Access, 2020, 8: 175840-175858.

[4] Byon Y J, Liang S. Real-time transportation mode detection using smartphones and artificial neural networks: Performance comparisons between smartphones and conventional global positioning system sensors. Journal of Intelligent Transportation Systems, 2014, 18 (3): 264-272.

[5] Semi-engineering. Retrieved from: https://semiengineering.com/what-does-an-ai-chip-look-like/.

[6] Momose H, Kaneko T, Asai T. Systems and circuits for AI chips and their trends. Japanese Journal of Applied Physics, 2020, 59 (5): 050502.

[7] Merolla P A, Arthur J V, Alvarez-Icaza R, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science, 2014, 345 (6197): 668-673.

[8] Painkras E, Plana L A, Garside J, et al. SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation. IEEE Journal of Solid-State Circuits, 2013, 48 (8): 1943-1953.

[9] Abderrahmane N, Lemaire E, Miramond B. Design space exploration of hardware spiking neurons for embedded artificial intelligence. Neural Networks, 2020, 121: 366-386.

[10] Han S, Mao H, Dally W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv: 1510.00149, 2015.

[11] Akopyan F, Sawada J, Cassidy A, et al. Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip. IEEE transactions on computer-aided design of integrated circuits and systems, 2015, 34 (10): 1537-1557.

[12] Sze V, Chen Y H, Yang T J, Emer J S. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proceedings of the IEEE, 2017, 105 (12): 2295-2329, Dec.

[13] Liu L, Zhao J, Cao G, et al. A Memristor-Based Silicon Carbide for Artificial Nociceptor and Neuromorphic Computing. Advanced Materials Technologies, 2021, 6 (12): 2100373.

[14] Huang B, Li N, Wang Q, et al. Optoelectronic Synapses Based on MoS2 Transistors for Accurate Image Recognition. Advanced Materials Interfaces, 2022, 9 (32): 2201558.

[15] Chen S, Li Z, Zhang S, et al. Emerging biotechnology applications in natural product and synthetic pharmaceutical analyses. Acta Pharmaceutica Sinica B, 2022.

[16] Kim Y, Shin D, Lee J, et al. A 0.55 V 1.1mW artificial intelligence processor with on-chip PVT compensation for autonomous mobile robots. IEEE Transactions on Circuits and Systems I: Regular Papers, 2017, 65 (2): 567-580.