Application analysis and future development of AI hardware acceleration period

Jiaming Hu*

School of Advanced Technology, School of XianJiaotong Liverpool University Suzhou, 215123, China

* Corresponding Author Email: Jiaming.Hu22@student.xjtlu.edu.cn

Abstract. AI hardware gas pedal is a hardware device specialized in accelerating the processing of Artificial Intelligence (AI) tasks, and its development history is closely related to the development of AI field. This will explore the theoretical foundation, development history, structural principles, and characteristics of AI hardware gas pedals. The theoretical foundations stem from the significant demands of deep learning algorithms for large-scale matrix operations, where traditional CPUs and GPUs encounter performance limitations. The development history of these accelerators shows a progression from general-purpose processors to the adoption of GPUs, which initially offered substantial improvements. However, as neural network models grew in complexity, GPUs too began to show bottlenecks. This led to the creation of specialized AI hardware accelerators which, through continual innovation, have become integral to AI computing. These devices are characterized by their efficient performance, low power consumption, and adaptability, offering a higher energy-efficiency ratio and faster processing for tasks like inference and training compared to conventional processors. Moreover, their architecture is optimized with customized computing cores, specialized storage systems, and effective data pathways to enhance the handling of extensive neural network models.

Keywords: AI hardware accelerator, gas pedal, advantages and disadvantages, development prospect.

1. Introduction

The surge in interest and applications of artificial intelligence across various sectors has necessitated the development of more efficient computational technologies. As AI models, particularly those based on deep learning, become increasingly complex, the need for specialized hardware to handle their computational demands has intensified. Traditional computing systems such as CPUs and GPUs, while foundational, struggle to meet the processing speeds and energy efficiency required for modern AI tasks. This challenge has spurred the evolution of AI hardware accelerators, which are specifically engineered to boost the performance of neural network computations, thereby enabling faster development and deployment of AI technologies in real-world applications.

First of all, the theoretical foundation of AI hardware gas pedals can be traced back to the development of the artificial intelligence field. With the rise of deep learning algorithms and neural network models, the demand for large-scale matrix operations and vector computations has increased dramatically, while it occurs problem for traditional CPUs (central processing units) and GPUs (graphics processing units) to process the same kind of issue. In order to perform these tasks more efficiently, the development of hardware gas pedals specialized for accelerating neural network inference and training has been explored. Second, the development history of AI hardware gas pedals has gone through several stages. Initially, people tried to use general-purpose processors (e.g., CPUs) for neural network computation, but they were not efficient due to their computational capacity limitations. Subsequently, GPUs were introduced into neural network computation due to their powerful performance in graphics processing, and a major breakthrough was achieved. However, with the increasing complexity of neural network models, GPUs are gradually showing performance bottlenecks. To address this challenge, AI hardware gas pedal emerged as a hardware device specialized in neural network computing, and through continuous innovation and iteration, it gradually became an important part of AI computing. The architectural principles of AI hardware gas

pedals mainly include components such as processing units, storage systems, and data pathways. As a hardware device specifically designed for neural network computation, the processing unit therein usually adopts a customized computing core that can efficiently perform neural network computations such as matrix multiplication and convolution. Meanwhile, the design of the storage system is also optimized for the characteristics of neural network models to improve the efficiency of data reading and transmission. In addition, for accelerating the process of computation of models that have large-scale neural networks, the design of data pathways is also crucial to manage the data flow and computation process efficiently. Finally, AI hardware gas pedals are characterized by efficient performance, low power consumption, and customizability. Compared to traditional general-purpose processors and graphics processors, AI hardware gas pedals have a higher energy-efficiency ratio when dealing with neural network computation tasks, and are able to complete tasks such as inference and training more quickly while reducing energy consumption. In addition, AI hardware gas pedals are often designed with the needs of different application scenarios in mind, and are able to provide flexible customization options to meet AI tasks of varying scale and complexity.

2. Case study and prospect of AI hardware accelerator

2.1. Case study

First of all, after understanding the AI hardware gas pedal we need to have a basic knowledge of it, such as the AI hardware gas pedal is what it can bring us what kind of help and what it is made up of things. The first thing I'd like to introduce is one of the AI hardware gas pedals, which we call a heterogeneous system. Using accelerators, we can achieve much higher system performance by recasting the CPU as the conductor of an AI orchestra. The CPU now plays a critical role in feeding the GPUs with data, orchestrating the workload, running the serial parts of the code that do not run efficiently on the GPUs, and managing all of the other services it takes for the computer. Heterogeneous systems are characterized by a flexible assemblage of specialized components, as opposed to reliance on a single multi-purpose CPU [1].

After learning about heterogeneous systems, I'd like to go over a very important piece of the AI hardware gas pedal, which we call the software stack. One crucial component that needs to be considered in the software stack is the synchronization of data traffic and the deployment of a new virtual machine. This means that all exchanges that occur between the SW and the VM, including memory transactions for the accelerator itself and any monitoring data that may be transferred at any point, must be finished before the VM is retired and the programmable fabric is reconfigured. The SW stack takes special care of this by synchronizing the threads passing the data from the Open Stack agent on to the fabric; if stale data remain in the AXI interconnection, this will lead to unpredictable behavior from which it may be will become a problem to recover. Consequently, when a fresh deployment request is prepared for execution, the software will be alert, at which point it ceases accepting new transfers and locks the deployment thread until all pending transactions have been finished. After the reconfiguration is completed, the locks are released and traffic to and from the accelerator can resume. The VM image, or more accurately the bitstream, can then be passed on to the DevC driver [2].

After understanding the above things, we will next introduce the knowledge related to combining edge AI with data. Placing computing power close to the source of data is the idea of an edge device, also similar to the idea of placing it close to the desktop or rack computer. Nevertheless, because of their size and power consumption, which are major drawbacks in many environments, specific Edge AI devices have been designed. The AI accelerator modules are the result of the need for specific mathematical capabilities, and their size and wireless connectivity make them easily attachable to industrial environments. Limited power consumption is also essential when many devices are deployed at once [3]. We can learn from this that there are still some shortcomings in Edge Accelerator, so what can be done to optimize these aspects. We have presented our ideas and methodology for the automated construction of edge AI computing platforms with customized

hardware accelerators, and although the scope of the framework we have presented is currently limited, we believe it can already significantly reduce the integration overhead for hardware developers. Hardware generation languages enable us to leverage concepts common to the software development domain, like APIs, which we believe can significantly accelerate hardware development cycles. Traditional methods, like graphical IP composers or HLS, tend to have limited design flexibility and require manual interface annotations. We plan to extend the existing generator API, for example with provisions for static verification; hardware generation languages are arguably even more flexible than traditional hardware description languages, and generator APIs offer a natural way of combining the RTL description with interface semantics. While this can help developers identify interface specification violations, it can also be used by the generation framework to optimize the generated hardware by using statically derived properties [4].

2.2. Advantages and disadvantages prospect

In the area of AI, AI hardware gas pedals are one of the key technologies, have both advantages and disadvantages in accelerating neural network computation. This paragraph will evaluate and compare the advantages and disadvantages of AI hardware gas pedals.

First, one of the advantages of AI hardware gas pedals lies in their efficient computational performance. Because they are specifically designed for neural network computation, AI hardware gas pedals are typically able to perform computational tasks such as matrix multiplication, convolution, and other computational tasks with higher computational efficiency and throughput than general-purpose processors (e.g., CPUs) and graphics processors (e.g., GPUs). This advantage enables AI hardware gas pedals to significantly reduce processing time and improve computational efficiency in large-scale neural network training and inference tasks. Second, AI hardware gas pedals typically have low power consumption. Designed with hardware architectures specifically for neural network computation, AI hardware gas pedals are able to utilize energy resources more efficiently when performing computational tasks, and typically have lower power consumption compared to general-purpose processors and graphics processors. AI hardware gas pedals process is shown in figure 1 [5].

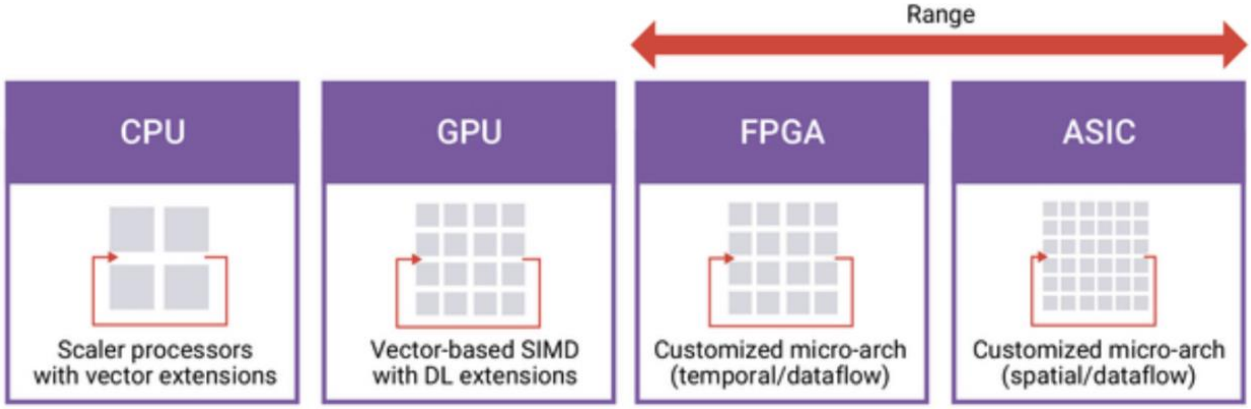


Fig. 1 AI hardware gas pedals process [5]

This feature enables AI hardware gas pedals to better meet the requirements of energy saving and emission reduction in large-scale AI application scenarios and reduce computing costs. In addition, AI hardware gas pedals have good customizability. For different neural network models and application scenarios, AI hardware gas pedals are usually able to provide customized design options to meet the needs of different users. This enables AI hardware gas pedals to provide more flexible and efficient solutions in a variety of complex AI tasks. However, corresponding to its advantages are some disadvantages. First, AI hardware gas pedals are expensive to design and develop. Due to the specialized design of AI hardware gas pedals, their R&D and production processes usually require more costly investment, which may increase the procurement cost of the hardware and limit its use by some small businesses or individual users. Second, AI hardware gas pedals are less compatible

and versatile. Because they are designed specifically for neural network computing, AI hardware gas pedals may not perform well when dealing with other types of computing tasks, which makes their application in some general-purpose computing scenarios somewhat limited. In addition, the update cycle of AI hardware gas pedals is short. With the continuous development of AI technology, new neural network models and algorithms continue to emerge, leading to higher requirements for the performance and functionality of AI hardware gas pedals. As a result, to follow the development of the technology, users may need to update their hardware more frequently, increasing the cost of maintenance and updates. In summary, AI hardware gas pedals have a series of advantages such as efficient computational performance, low power consumption, and good customizability, but also face disadvantages such as higher design costs, poorer compatibility, and shorter update cycles. In practical applications, users need to comprehensively consider various factors and choose the AI hardware gas pedal solution that best suits their needs.

As a result, one of the main focuses of research in academia and industry has been to specialize computing systems for AI. Unlike traditional system design, designers of AI-specialized platforms must balance two seemingly incompatible requirements. AI systems should, on the one hand, provide extremely high compute efficiencies in order to support their deployment in scenarios with both high throughput (such as data centers and the cloud) and low energy consumption (such as edge devices). However, in order for AI systems to be programmed to fully utilize their potential, the end-user should have to put in little to no extra effort. Meeting these needs democratizes AI systems for non-expert users, which is essential to their widespread adoption [5].

Although AI has been gradually popularized in human life, and its help to human beings has made great progress compared to before, but relatively, it still has many problems and will face challenges in the future. The first is about the problem of energy consumption, in the operation of the AI hardware gas pedal, due to the need for AI to carry out a very huge computing, but its energy consumption is also a very headache, and secondly, AI does not always play a positive role in people's lives to help, in some aspects of the AI cannot give us a comprehensive help, and even some times will give us the wrong answer, so I think the In the future development of AI hardware gas pedal, people should pay attention to this problem, so that AI can give more accurate and helpful responses, so in the next R & D, I think the development of AI is still facing a lot of problems, because AI is a field that human beings have never been involved in, so it can be said that this is a very great challenge for all of mankind.

3. Challenges and prospects

The future of AI hardware accelerators, often referred to as "AI hardware gas pedals," is poised for transformative growth and innovation, driven by the escalating demands of AI applications for faster computation, greater efficiency, and lower energy consumption. As AI models become increasingly complex, traditional CPU-based computing platforms are reaching their limits, necessitating the development of specialized AI accelerators. This chapter describes how machine learning applications can run more efficiently when hardware acceleration is used [6].

In the coming years, we can expect to see a significant evolution in AI hardware accelerators, with a focus on three key areas: performance, energy efficiency, and adaptability. Performance-wise, future accelerators will leverage advancements in semiconductor technology, such as the adoption of smaller nanometer fabrication processes, to increase computational power while reducing physical size. This will enable the deployment of AI capabilities in a wider range of devices, through side computing devices to large centers of data, intensifying the ubiquity of Artificial Intelligence in everyday applications.

Through the impact of the environment on computing increasingly scrutinized, more attention will be put on the efficiency of energy. Next-generation AI accelerators will incorporate novel architectures and materials, such as neuromorphic computing elements and low-power, high-performance semiconductors, to drastically reduce power consumption. This will be crucial for

enabling sustainable AI operations particularly in large-scale deployments. Nevertheless, up until now, the inadequate performance of inference accelerators has kept input image frame rates and resolutions during real-time inference limited [7].

Adaptability will also be a hallmark of future AI accelerators. As AI applications diversify, there will be a growing need for hardware that can support a wide range of AI models and algorithms. Future accelerators will likely feature more programmable and reconfigurable elements, allowing them to be optimized for specific tasks or to adapt to new AI paradigms without requiring physical replacement. It offers information on the most recent developments and trends in neural network hardware accelerators [8].

Moreover, the integration of AI accelerators together along the company's other appearing technologies, including quantum computing and advanced networking, will open up new possibilities for AI applications, from ultra-precise simulations to actual-time, intelligent data processing at the network's rim. Three design challenges have been addressed with a variety of optimization techniques [9].

In summary, the future of AI hardware accelerators is bright, with innovations in performance, efficiency, and adaptability set to redefine the landscape of computing. However, their computational demands make them challenging for general-purpose architectures to handle [10]. These improvements contribute to more sustainable, efficient, and versatile AI deployments across a myriad of industries rather than just improving the capabilities of AI systems.

4. Conclusion

In conclusion, the landscape of AI hardware accelerators is undergoing rapid transformation, spurred by the growing computational demands of advanced AI applications. These accelerators are increasingly central to achieving higher computational efficiencies, particularly in scenarios requiring extensive data processing like cloud computing and edge devices. The advantages of AI hardware accelerators are marked by their ability to perform specialized tasks with greater speed and reduced energy consumption, making them integral in driving forward the capabilities of neural network computations and other AI-driven technologies.

However, these benefits are tempered by significant challenges. The specialized nature of AI accelerators leads to high design and development costs, and their focus on specific computational tasks can limit their versatility and necessitate frequent updates to keep pace with advancing AI technologies. Additionally, issues like energy consumption and the need for tailored programming remain critical hurdles.

Looking ahead, the future of AI hardware accelerators is poised for innovative growth. Developments in semiconductor technology and new architectures like neuromorphic computing are expected to further enhance performance and energy efficiency. Furthermore, the adaptability of these systems is set to improve, allowing them to support a broader array of AI models and applications without the need for physical replacements.

References

[1] Welser J, Pitera J W, Goldberg C. Future computing hardware for AI. 2018 IEEE International Electron Devices Meeting (IEDM). IEEE, 2018: 1.3. 1-1.3. 6.

[2] Karras K, Pallis E, Mastorakis G, et al. A hardware acceleration platform for AI-based inference at the edge. Circuits, Systems, and Signal Processing, 2020, 39(2): 1059-1070.

[3] Sipola T, Alatalo J, Kokkonen T, et al. Artificial intelligence in the IoT era: A review of edge AI hardware and software. 2022 31st Conference of Open Innovations Association (FRUCT). IEEE, 2022: 320-331.

[4] Hielscher L, Bloeck A, Viehl A, et al. Platform generation for edge AI devices with custom hardware accelerators. 2021 IEEE 19th International Conference on Industrial Informatics (INDIN). IEEE, 2021: 1-8.

[5] Venkataramani S, Choi J, Srinivasan V, et al. Deeptools: Compiler and execution runtime extensions for rapid ai accelerator. IEEE Micro, 2019, 39(5): 102-111.

[6] Talib M A, Majzoub S, Nasir Q, et al. A systematic literature review on hardware implementation of artificial intelligence algorithms. The Journal of Supercomputing, 2021, 77(2): 1897-1938.

[7] Park H, Kim S. Hardware accelerator systems for artificial intelligence and machine learning. Advances in Computers. Elsevier, 2021, 122: 51-95.

[8] Mohaidat T, Khalil K. A Survey on Neural Network Hardware Accelerators. IEEE Transactions on Artificial Intelligence, 2024.

[9] 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.

[10] Dhilleswararao P, Boppu S, Manikandan M S, et al. Efficient hardware architectures for accelerating deep neural networks: Survey. IEEE access, 2022, 10: 131788-131828.