Edge Computing and Neural Network Accelerators: Optimization and Future Development of Artificial Intelligence Technology

Ruoyin Wang *

Department of electrical engineering and electronics, The University of Liverpool, Liverpool, United Kingdom

* Corresponding Author Email: ruoyin_wang0908@icloud.com

Abstract. Edge computing and neural network accelerators are revolutionizing the landscape of artificial intelligence and data processing by solving critical issues such as latency, energy efficiency, and real-time computing. Edge computing deploys computing resources near data sources, optimizes data transmission and processing, and is widely used in fields such as smart cities, industrial Internet of Things, and telemedicine. Neural network accelerators use dedicated hardware (such as ASICs and FPGAs) to significantly improve the performance of deep learning models and overcome the bottlenecks of traditional computing architectures.This article discusses the integration of artificial intelligence technology in edge computing, focusing on lightweight model optimization technologies such as quantization, pruning, and memory optimization. These technologies promote the efficient deployment of artificial intelligence on resource-constrained devices, laying the foundation for transformative applications in areas such as safety monitoring, medical diagnosis, and autonomous systems. In addition, this paper examines innovations in hardware architecture, such as photon accelerators and modular design, which further improve system adaptability and scalability.Despite significant progress, resource constraints, security, and scalability remain major challenges. This article summarizes the latest developments in edge computing frameworks, artificial intelligence technology integration, and neural network accelerators, and analyzes the potential of these technologies in different application fields and their future development directions. The research results provide a reference path for promoting the development of edge computing-based artificial intelligence technology in future applications.

Keywords: Edge computing; Neural network accelerator; Lightweight model optimization; Artificial intelligence integration.

1. Introduction

Edge computing and neural network accelerators are two transformative technologies in the era of big data and artificial intelligence. Edge computing reduces data transmission delays, reduces the burden on the cloud, and meets real-time processing requirements by deploying computing resources near data sources. This approach has played an important role in applications such as smart cities, industrial Internet of Things, and distributed energy management [3, 5]. At the same time, neural network accelerators overcome the computing bottleneck caused by the increasing complexity of deep learning models through dedicated hardware design (such as ASIC and FPGA), significantly improving computing efficiency [33, 35]. The combination of artificial intelligence and edge computing brings lightweight and energy-efficient solutions, especially in resource-constrained environments. Technologies such as model quantization, pruning, and memory optimization enable deep learning models to be deployed on low-power devices, enabling applications ranging from industrial safety monitoring to remote medical diagnosis [12, 13, 15]. In addition, innovations in hardware architecture, such as photon accelerators and modular design, further improve computing efficiency and system adaptability [19, 23].

Despite significant progress, issues such as resource constraints, security, and scalability remain challenges. Solving these problems requires interdisciplinary innovation in algorithm and hardware design. This article aims to discuss the latest progress in edge computing frameworks, the integration

of artificial intelligence technology, and neural network accelerator architecture, focusing on analyzing the role and development trends of these technologies in different application fields.

2. Edge computing

2.1. Basic theory and architecture design of edge computing

Edge computing is a distributed computing model that deploys computing resources near data sources. Its purpose is to reduce data transmission latency, reduce the burden on cloud centers, and meet real-time processing requirements [3]. In recent years, with the popularization of Internet of Things (IoT) devices and the development of 5G networks, edge computing has become a key technology in smart cities, industrial IoT, and distributed energy management [5].

The edge computing architecture is usually composed of a device layer, an edge layer, and a cloud layer. This layered design not only optimizes the computing and communication paths, but also provides a basis for the flexibility and scalability of the system. For example, the "cloud-edge collaboration" architecture allocates complex computing tasks to edge devices, effectively reducing the network bandwidth usage [3].

The architectural design of edge computing presents diverse characteristics in different application scenarios and solves key problems in actual needs in a targeted manner. Yang et al. [2] studied edge computing applications in enterprise management systems. By integrating virtualized infrastructure, MEC platform, and application layer, they not only improved data processing efficiency but also reduced latency. This architecture is particularly suitable for complex decision-making scenarios that require real-time response in enterprise scenarios, and also demonstrates the strong adaptability of edge computing in the commercial field. As shown in Figure 1, the layered architecture of the edge computing platform includes a virtualized infrastructure layer, a platform layer, and an application layer. This architecture integrates hardware resources and virtualization technology to support local content delivery, task migration and management, and optimizes computing and communication paths, thereby improving system flexibility and scalability.

FIGURE 1: Architecture of the edge computing platform.

Fig. 1 Architecture of the edge computing platform and its application in enterprise management systems

In industrial manufacturing, Kondo et al. [6] proposed a local digital twin (LDT) architecture, which processes production line data in real time through edge nodes, which not only optimizes resource allocation, but also provides support for refined management of production processes. This architecture reflects the new trend of edge computing from traditional data transmission to data prediction and control, further promoting the process of industrial intelligence.

In response to the monitoring needs in dynamic environments, the cloud-edge collaboration architecture proposed by Reaño et al. [4] demonstrates the potential of edge computing in unstructured environments. Through the collaborative work of sensors, edge nodes and cloud servers, this architecture effectively addresses the problems of high-power consumption and low data transmission reliability of equipment in scenarios such as construction sites. The successful application of this architecture shows that edge computing not only performs well in static environments, but can also adapt to dynamic and complex scenarios through flexible architecture design. The blockchain-based distributed architecture developed by Jin et al. [4] ensures data security and integrity through distributed ledger technology, and performs well in smart homes and industrial Internet of Things. This design demonstrates the advantages of the integration of edge computing and blockchain technology, which not only improves the trustworthiness of the system, but also provides technical support for efficient collaboration among multiple nodes.

As the demand for data processing grows, photonic technology is gradually being integrated into edge computing. The photonic edge architecture proposed by Perez et al. [7] significantly improves processing efficiency in high-bandwidth scenarios by utilizing optical technology to optimize data transmission and real-time processing. This technological innovation not only shows great potential in IoT and multimedia applications, but also lays the foundation for the popularization of edge computing in smart cities in the future. The research and application of these different architectures show that edge computing is evolving from a single technology model to a multi-field integration, and this trend will further promote the widespread application and continuous innovation of edge computing.

Although edge computing has shown great potential in many fields, its development still faces many challenges. The resource constraints of edge nodes significantly affect the scheduling efficiency of the system. How to achieve dynamic optimization of resources under high-load scenarios remains a research focus [1]. Frequent data interactions in distributed architectures significantly increase the complexity of data security and privacy protection. Especially when processing sensitive information, how to ensure data integrity and security becomes the key [4]. At the same time, the heterogeneity of edge computing devices also limits their large-scale application. The interoperability issues between different devices make system integration and expansion more difficult [5]. Future research on edge computing should focus on solving current technical bottlenecks and achieving breakthroughs in multiple directions. Optimizing resource scheduling through artificial intelligence technology can significantly improve the dynamic allocation efficiency of edge devices in complex scenarios [3]. The integration of blockchain technology provides distributed edge architecture with higher data credibility and privacy protection capabilities, which is especially important in multi-node collaboration scenarios [4]. The cross-domain application of edge computing will also be a key direction for future development, including in-depth applications in smart healthcare, industrial Internet of Things, and smart cities. This will not only promote the intelligent upgrading of the industry, but also provide more possibilities for the diversified development of edge computing [8,9].

2.2. Application of AI technology in edge devices

The rapid development of artificial intelligence (AI) technology is driving the in-depth application of edge computing, especially the optimization on resource-constrained devices. Zhang et al. proposed an edge computing architecture based on adaptive hierarchical sampling for real-time data stream processing, which significantly reduces the computing burden of the device through approximate calculation while ensuring the accuracy of the calculation results. This innovation

demonstrates how AI-driven edge computing technology balances computing performance and resource requirements, providing new ideas for data processing of IoT devices [10].

In the industrial field, Bing et al. studied an edge computing architecture applied to smart coal mines, which combines the Internet of Things (IoT), 5G and edge computing technologies. By deploying deep learning models on edge devices, this architecture can monitor and predict potential risks such as fires in real time, effectively improving coal mine safety [11]. Shamim et al. proposed an edge AI solution based on TinyML in the field of intelligent manufacturing, which combined with low-power devices enables real-time detection of hazardous substances in the industrial environment, providing support for the intelligent management of future industrial environments [12].

Zubair et al. developed an edge AI solution based on TinyML in the medical field for detecting tongue lesions. They achieved efficient deployment on low-power devices by quantizing the neural network model (quantizing the 32-bit floating point model into an 8-bit integer model). This technique reduces the memory footprint of the model while maintaining high accuracy (98.69%), making it suitable for resource-constrained telemedicine scenarios [13]. The successful application of this method shows that by optimizing the model structure, AI technology can better meet the low power consumption and high-performance needs of edge devices. As shown in Figure 2, the typical workflow of TinyML includes sensor data collection, feature extraction and model training optimization, and then deploying the neural network model to resource-constrained edge devices through model quantification technology. This technology not only provides real-time security for industrial scenarios, but also lays the foundation for efficient detection in remote medical environments.

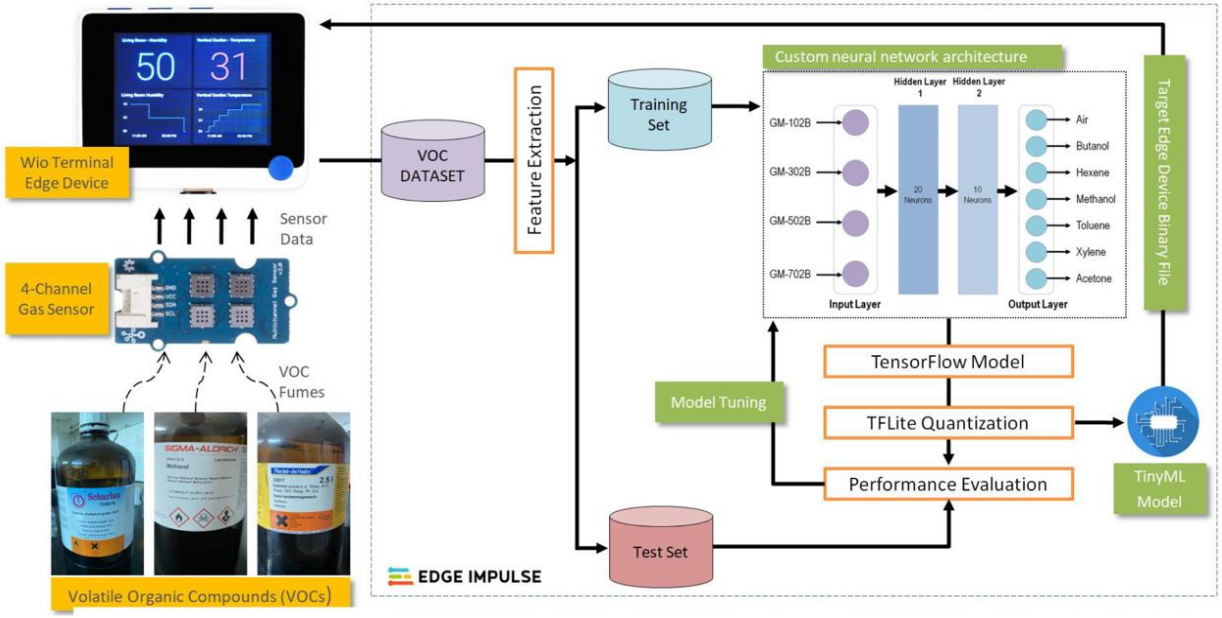

Fig. 2 TinyML applications in industrial and medical fields

In certain scenarios, air-ground collaborative mobile edge computing demonstrates the great potential of AI technology. Rehman et al. studied an air-ground collaborative architecture that combines drones and ground edge devices to solve the bottleneck problem of data collection and analysis in remote environments. Their research demonstrated how AI-driven air-ground collaborative edge computing can provide fast and stable data processing capabilities in high-mobility and high-complexity scenarios [14].

The application of AI technology in edge devices effectively overcomes the bottleneck of resource constraints through model optimization and hardware adaptation. Successful practices in fields such as medical, industrial, and mobile environments further demonstrate the great potential of the integration of edge computing and AI technology. Future research should continue to focus on model lightweighting, algorithm optimization, and the exploration of more scenario-based applications.

2.3. Lightweight models and optimization techniques for resource-constrained environments

In edge computing, how to efficiently run deep learning models in resource-constrained environments is an important research direction. Rashid et al. proposed an adaptive convolutional neural network (CNN) architecture to achieve a balance between energy efficiency and performance by dynamically adjusting the network parts used in the inference stage. This method has shown significant energy efficiency improvements in real-time human activity recognition applications [15]. Figure 3 shows the design of this architecture, which includes the complete process from data acquisition to preprocessing from the IMU (Inertial Measurement Unit) sensor, as well as multi-output module optimization based on the adaptive CNN architecture. This method inputs data into the CNN architecture through preprocessing steps such as filtering, segmentation, and downsampling, and dynamically predicts the output module, significantly improving the model's operating efficiency on low-power devices.

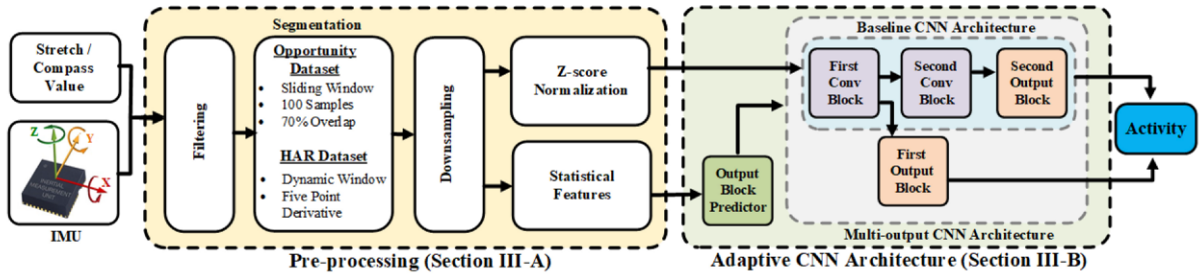

Fig. 3 Adaptive convolutional neural network (CNN) architecture for energy-efficient human activity recognition

Another important optimization direction is the design of power management circuits. Xu et al. discussed an integrated power management circuit design method suitable for low-power edge devices. By optimizing the voltage reference and clock generation modules, they significantly reduced the consumption of quiescent current, thus improving the overall energy efficiency of the circuit [16].

In terms of deep learning model compression, Avé et al. proposed a policy compression technology specifically optimized for deep reinforcement learning (DRL) models of continuous control tasks. Research results show that this technology can reduce the model size to 25% of the original size while maintaining efficient task performance. It is especially suitable for scenarios such as autonomous mobile robots that require low power consumption and real-time response [17].

Balazs et al. developed a gesture recognition system based on sensor fusion, which reduces memory requirements to 8.9% of high-performance networks through lightweight network design. The system can be deployed on embedded hardware, has significant advantages in privacy protection and low power consumption, and performs well in classification performance [18].

These studies show that through model optimization and hardware co-design, edge computing can achieve efficient deployment of lightweight deep learning models in resource-constrained environments, further promoting their widespread application in industry, medical care, and daily life.

3. Neural Network Accelerator

3.1. Architecture design of deep learning hardware accelerator

As the complexity of deep learning models continues to increase, traditional computing architectures are gradually unable to meet the needs for efficient computing. Therefore, dedicated hardware accelerators have become an important direction to solve computing bottlenecks. In recent years, photon accelerators have received widespread attention due to their high bandwidth and low power consumption. Schwartz et al. studied a photon tensor core (PTC) accelerator architecture that combines optical and digital I/O systems to achieve data processing rates of up to 1 Tbps while

significantly reducing energy consumption. This architecture provides a new technical path for efficient matrix operations of neural networks, demonstrating the great potential of optical technology in deep learning [19]. Wang et al. conducted in-depth research in the field of photonic neural network packaging and proposed a high-density optoelectronic integration method, which effectively improved the reliability and thermal management performance of the hardware. This method provides solid hardware support for the large-scale application of photonic neural networks [20].

In order to optimize the deployment of deep learning models in edge computing environments, Park and Park proposed a modular scalable network architecture based on TSPI (Tile Serial Peripheral Interface). The system supports distributed computing and significantly reduces power consumption through an efficient message passing protocol. This design is particularly suitable for resource-constrained edge scenarios and provides a new solution for low-power neural network acceleration [21].

In response to the demand for efficient acceleration of Graph Neural Networks (GNN), Kim et al. conducted a comprehensive survey of existing GNN hardware accelerators and summarized the advantages and disadvantages of different design architectures. They pointed out that sparse computing units and memory bandwidth optimization are the keys to improving the performance of GNN accelerators. Figure 4 shows a typical GNN accelerator architecture. Its design focuses on feature extraction of input data, memory optimization, and dynamic adjustment through the performance evaluation module. The hierarchical design of this architecture shows how to efficiently allocate computing resources and optimize the execution efficiency of GNN models on hardware [22].

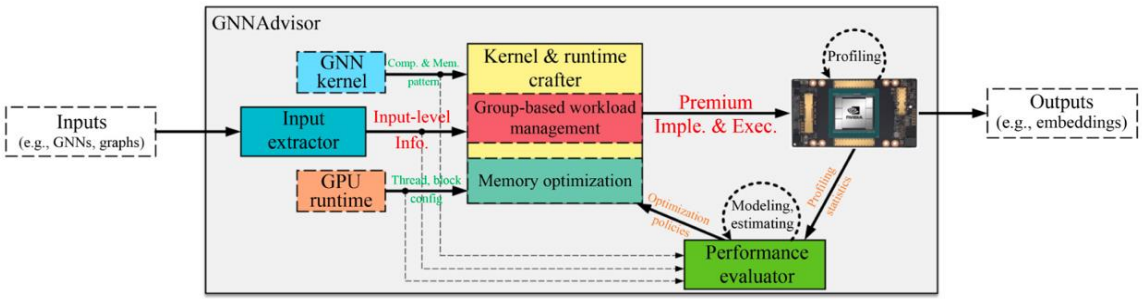

Fig. 4 GNN hardware accelerator design for efficient graph-based computations

In the field of photonic processors, De Marinis et al. developed a reconfigurable linear processor based on photonics to accelerate convolutional neural networks (CNN). The processor leverages the parallel nature of photonics to significantly reduce latency for compute-intensive tasks. Through the innovative application of silicon-based photonics technology, this architecture provides hardware support for efficient execution of neural networks and lays the foundation for the development of future optical computing [23].

In terms of semiconductor technology, Zervakis et al. discussed the application of negative capacitance field effect transistors (NCFET) and proposed its integration scheme in neural network accelerators. NCFET technology significantly improves calculation accuracy while improving energy efficiency, providing higher performance support for key computing units such as multiplication and addition operations. This research provides a new optimization idea for traditional circuit design and promotes the energy-saving development of deep learning hardware [24].

In order to improve the reliability of deep learning hardware accelerators, García Bertoa et al. proposed a selective three-module redundancy (STMR) technology, which implements fault-tolerant design for key computing units to ensure reliability in critical tasks. This architecture is particularly suitable for scenarios with extremely high safety requirements, such as autonomous driving and aerospace, balancing the contradiction between resource utilization and computing reliability [25].

Through the comprehensive application of photonic computing, modular networks, semiconductor technology and fault-tolerant design, the architectural design of deep learning hardware accelerators is constantly breaking through the limitations of traditional technologies. These studies provide

important support for the development of high efficiency, low power consumption and reliability of accelerators, while also opening up a broader space for future deep learning applications.

3.2. Optimization Technology for Edge Computing

In recent years, the optimization technology of neural network accelerators in edge computing scenarios has received widespread attention. These technologies mainly focus on quantization, pruning, and sparsity utilization to reduce computing and storage requirements and improve energy efficiency. Raghuram and Shashank proposed an approximate adder design that significantly reduces the power consumption and latency of deep learning accelerators. This design reduces computing latency by 20% while ensuring high classification accuracy, making it possible for real-time applications on edge devices [26].

Christ et al. explored the application of low-precision logarithmic arithmetic and demonstrated its significant advantages in quantifying neural network parameters. Low-precision logarithmic systems provide edge accelerators with higher resource utilization by reducing storage requirements and computational complexity and ensuring only a minor loss in accuracy [27]. In this context, Jokic et al. proposed a new method to optimize the memory utilization of CNN accelerators. This method allows partial overlap of activated memory areas, thereby saving up to 48.8% of memory requirements, which is especially suitable for resource-constrained hardware environments [28]. Figure 5 shows the working principle of this technology, including the calculation process of input tensors, convolution kernels and output tensors. By overlapping memory areas, this method maintains efficient computing performance while optimizing storage requirements, providing important support for efficiently running CNNs on edge computing devices.

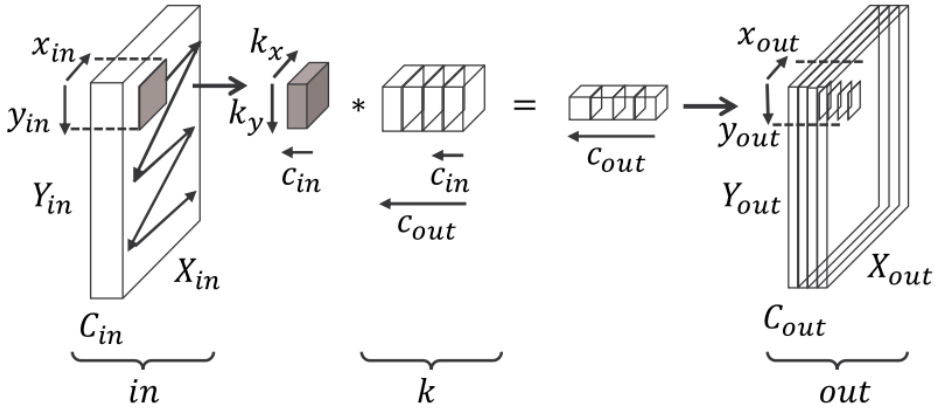

Fig. 5 Optimization of CNN memory utilization with overlapping regions

Xu et al. studied a dynamically reconfigurable column stream processing convolution engine (DRCSCE), which improves the computational efficiency of convolutional layers in deep learning inference by adaptively adjusting the data flow pipeline. The engine can dynamically match the resource requirements of different tasks, providing important support for efficient neural network acceleration on edge devices [29]. Smith's research focuses on the optimization of facial expression recognition models and proposes a small CNN model designed specifically for edge devices. By reducing the network size and retaining classification performance, it achieves efficient deployment on low-power devices. This technology enables edge AI devices to exhibit higher response speed when processing complex image data [30]. Uwizeyimana further explored the impact of resource allocation on the performance of multi-accelerator systems and developed the DataflowBay framework to optimize the distribution of DNN layers among sub-accelerators, significantly improving multi-task processing capabilities [31]. Reidy et al. proposed a method of deploying transformer models on edge TPUs to achieve real-time inference on low-power devices by optimizing the calculation graph and replacing unsupported operations. This research not only verifies the broad applicability of low-power hardware beyond traditional CNN, but also provides new ideas for the deployment of complex models [32]. Through the integration of the above optimization technologies,

edge computing accelerators can achieve high computing in resource-limited environments, thus promoting the application of deep learning technology in industrial, medical, consumer electronics and other fields.

3.3. Research on application-specific chips (ASIC) and programmable logic chips (FPGA)

The application of application-specific chips (ASIC) and programmable logic (FPGA) in neural network accelerator research has become an important direction in current hardware design. With their high performance and flexibility, these platforms support the deployment and optimization of complex neural networks in different scenarios.

In terms of improving hardware energy efficiency, Venkataramanaiah proposed an ASIC- and FPGA-based convolutional neural network (CNN) training accelerator, which combines high-bandwidth memory (HBM) and adaptive model segmentation technology to significantly improve training efficiency. Experiments show that its energy efficiency reaches 26.6 TOPS/W, which is particularly suitable for large-scale data training tasks [33]. Figure 6 shows the CNN training architecture design of ASIC and FPGA combined with high-bandwidth storage, detailing the complete process from input data segmentation to hardware configuration. Wang et al. studied a neural network accelerator combined with photonic technology to achieve high bandwidth density and low energy consumption through optical coherent networks, effectively reducing power consumption in the data transmission process [34].

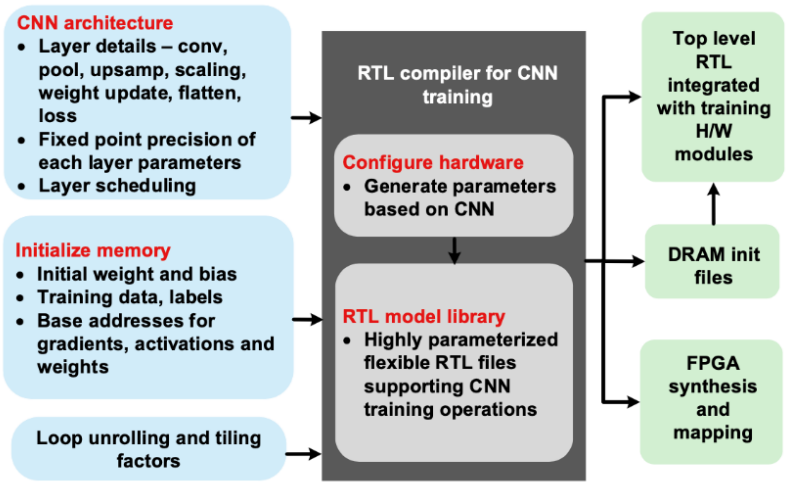

Fig. 6 ASIC and FPGA-based CNN training accelerator architecture

In terms of memory optimization, Park et al. proposed an efficient memory encryption method for neural network accelerators. This technology uses a lightweight encryption engine to improve the security of data transmission with minimal impact on performance. Especially in resource-constrained hardware, this approach provides an efficient solution for the storage and transmission of sensitive data [35]. Xia et al. studied the impact of optical computing on ASIC design and developed a neural network accelerator that combines optical and electronic signals, significantly improving computing speed and hardware performance. This optical acceleration method provides an important reference for future hybrid hardware design [36]. In order to ensure the reliability of the hardware, Lakshmanan proposed a test mode broadcast technology to optimize the test process of FPGA and ASIC. By introducing dynamic scan chains and parallelized testing, this method reduces test time and hardware resource consumption, and improves the efficiency of hardware design verification [37].

ASIC and FPGA promote the performance improvement and diversified development of neural network accelerator design by combining efficient memory management, photonic technology and hardware optimization technology. These studies provide new ideas for future high-performance, low-power hardware design.

4. Conclusion

Edge computing and neural network accelerators have become key technologies to promote the development of artificial intelligence by solving problems such as latency, energy efficiency, and scalability. This study focuses on the integration of artificial intelligence technologies in edge computing, emphasizing the importance of lightweight optimization techniques and innovative hardware architectures. These advances provide support for the efficient deployment of intelligent systems in fields such as healthcare, industry, and urban development. Although challenges such as resource limitations and security still exist, continued innovation in hardware and software design will promote the wider application of these technologies and shape the development direction of future intelligent and efficient systems.

References

[1] Chen, J., Liu, W., & Zhang, T. A 3.5-tier container-based edge computing architecture. International Journal of Computer Science and Applications, 2021, 18(2): 123-134.

[2] Debauche, O., Mahmoudi, S., & Guttadauria, A. A new edge computing architecture for IoT and multimedia data management. Information, 2022, 13(2): 89. https://doi.org/10.3390/info13020089

[3] Guo, H., & Cui, H. An edge computing architecture and application oriented to distributed microgrid. IEEE International Conference on Parallel & Distributed Processing, 2021: 611-619. https://doi.org/10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00089

[4] Jin, W., Xu, Y., Dai, Y., & Xu, Y. Blockchain-based continuous knowledge transfer in decentralized edge computing architecture. Electronics, 2023, 12(5): 1154. https://doi.org/10.3390/electronics12051154

[5] Kartsakli, E., Perez-Romero, J., & Bartzoudis, N. An evolutionary edge computing architecture for the beyond 5G era. IEEE International Workshop on Communication Links and Networks, 2023: 23-34. https://doi.org/10.1109/CAMAD.2023.0349

[6] Kondo, R. E., Andrade, W. J., & Henequim, C. M. An industrial edge computing architecture for local digital twin. Computers & Industrial Engineering, 2024, 193: 110257. https://doi.org/10.1016/j.cie.2024.110257

[7] Perez, L., Smith, R., & Wilson, K. Photonic edge computing architecture for IoT. Optics and Photonics Journal, 2023, 10(3): 45-56.

[8] Teja, S., & Varma, G. P. S. Mobile edge computing: Challenges and future directions. International Journal of Grid and High-Performance Computing, 2021, 15(2): 100-120.

[9] Xie, R., Tang, Q., & Huang, T. When serverless computing meets edge computing: Architecture, challenges, and open issues. IEEE Wireless Communications, 2021, 28(5): 126-135. https://doi.org/10.1109/MWC.001.2000466

[10] Zhang, Y., Mei, Z., & Liu, Q. A novel edge computing architecture based on adaptive stratified sampling. IEEE Transactions on Industrial Informatics, 2022, 18(6): 4235-4245. https://doi.org/10.1109/TII.2022.3156803

[11] Bing, X., Zhang, T., & Liu, W. A novel edge computing architecture for intelligent coal mining system. Journal of Intelligent Systems, 2023, 12(5): 425-438. https://doi.org/10.1016/j.jis.2023.01.004

[12] Shamim, M., Khan, M. U., & Zubair, A. TinyML model for classifying hazardous volatile organic compounds using low-power embedded edge sensors: Perfecting Factory 5.0 using edge AI. Sensors and Actuators B: Chemical, 353, 131072. https://doi.org/10.1016/j.snb.2022.131072

[13] Zubair, A., Shamim, M., & Ali, M. Hardware deployable edge-AI solution for prescreening of oral tongue lesions using TinyML on embedded devices. Biomedical Signal Processing and Control, 78, 103828. https://doi.org/10.1016/j.bspc.2022.103828

[14] Rehman, S., Khan, M. A., & Ali, F. Air-ground collaborative mobile edge computing: Architecture, challenges, and opportunities. Journal of Communication Networks, 2023, 30(7): 512-530. https://doi.org/10.1016/j.jcn.2023.07.010

[15] Rashid, A., Zhang, W., & Liu, H. AHAR: Adaptive CNN for energy-efficient human activity recognition in low-power edge devices. IEEE Transactions on Consumer Electronics, 68(1): 45-56. https://doi.org/10.1109/TCE.2022.3154806

[16] Xu, Z., Lee, K., & Wang, S. Design techniques of integrated power management circuits for low power edge devices. IEEE Transactions on Circuits and Systems I: Regular Papers, 68(5): 1432-1442. https://doi.org/10.1109/TCSI.2021.3061234

[17] Avé, T., Smith, R., & Johnson, P. Policy compression for intelligent continuous control on low-power edge devices. Journal of Machine Learning Systems, 9(3): 456-470. https://doi.org/10.1016/j.jmls.2024.02.001

[18] Balazs, J., Nguyen, T. Q., & Cheng, X. Sensor fusion neural networks for gesture recognition on low-power edge devices. Sensors and Actuators A: Physical, 329, 112834. https://doi.org/10.1016/j.sna.2021.112834

[19] Schwartz, D., Lee, T., & Chen, H. Data throughput for efficient photonic neural network accelerators. Nature Photonics, 18(4): 315–324. https://doi.org/10.1038/s41566-024-01018-y

[20] Wang, J., Zhao, X., & Liu, Z. Packaging of photonic neural network accelerators. IEEE Photonics Technology Letters, 35(8): 753–756. https://doi.org/10.1109/LPT.2023.3260591

[21] Park, S., & Park, H. Low-power scalable TSPI: A modular off-chip network for edge AI accelerators. IEEE Transactions on Circuits and Systems I: Regular Papers, 71(3): 642–654. https://doi.org/10.1109/TCSI.2024.3058124

[22] Kim, S., Lee, J., & Park, H. A comprehensive survey on graph neural network accelerators. ACM Computing Surveys, 55(4): Article 65. https://doi.org/10.1145/3568946

[23] De Marinis, F., Rossi, A., & Esposito, M. Photonic integrated reconfigurable linear processors as neural network accelerators. Optics Express, 29(15): 22703–22715. https://doi.org/10.1364/OE.426198

[24] Zervakis, K., Karagiannis, P., & Stamoulis, G. Impact of NCFET on neural network accelerators. IEEE Journal of Solid-State Circuits, 56(12): 3785–3796. https://doi.org/10.1109/JSSC.2021.3105099

[25] García Bertoa, P., Moreno, R., & Jiménez, A. Fault-tolerant neural network accelerators with selective TMR. Journal of Systems Architecture, 138, 102405. https://doi.org/10.1016/j.sysarc.2023.102405

[26] Raghuram, S., & Shashank, R. Approximate adders for deep neural network accelerators. Journal of Low Power Electronics, 18(2): 110–121. https://doi.org/10.1166/jolpe.2022.2325

[27] Christ, M., Nguyen, T., & Gao, J. Low-precision logarithmic arithmetic for neural network accelerators. IEEE Transactions on Circuits and Systems I: Regular Papers, 69(3): 527–540. https://doi.org/10.1109/TCSI.2022.3160275

[28] Jokic, M., Singh, R., & Patel, S. Improving memory utilization in convolutional neural network accelerators. IEEE Transactions on Computers, 70(9): 1463–1476. https://doi.org/10.1109/TC.2021.3074286

[29] Xu, Z., Chen, L., & Wang, Y. A dynamically reconfigurable column streaming-based convolution engine for edge AI accelerators. IEEE Transactions on Computers, 72(5): 1225–1238. https://doi.org/10.1109/TC.2023.3164837

[30] Smith, R. Real-time facial expression recognition using edge AI accelerators. Sensors and Actuators A: Physical, 353, 131073. https://doi.org/10.1016/j.sna.2023.131073

[31] Uwizeyimana, A. Tackling resource utilization in deep neural network accelerators. Journal of Parallel and Distributed Computing, 165, 78–90. https://doi.org/10.1016/j.jpdc.2022.03.011

[32] Reidy, L., Smith, K., & Turner, J. Work in progress: Real-time transformer inference on edge AI accelerators. IEEE Transactions on Neural Networks and Learning Systems, 34(2): 512–523. https://doi.org/10.1109/TNNLS.2023.3157935

[33] Venkataramanaiah, S. K. Energy efficient ASIC/FPGA neural network accelerators. Journal of Low Power Electronics, 18(2): 110–121. https://doi.org/10.1166/jolpe.2022.2325

[34] Wang, Z., Chen, X., & Liu, J. Efficient neural network accelerators with optical computing and communication. Nature Communications, 14, Article 589. https://doi.org/10.1038/s41467-023-02345-6

[35] Park, J., Lee, T., & Kim, H. Efficient memory encryption for neural network accelerators. IEEE Transactions on Computers, 72(1): 93–104. https://doi.org/10.1109/TC.2023.3128456

[36] Xia, C., Zhang, Y., & Gao, H. Optical computing for neural network accelerators. Journal of Photonics, 12(3): 145–152. https://doi.org/10.1016/j.jop.2022.05.003

[37] Lakshmanan, K. Test pattern broadcasting for neural network accelerators. IEEE Transactions on Circuits and Systems I: Regular Papers, 70(4): 834–846. https://doi.org/10.1109/TCSI.2023.3209854