Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD24520

Lightweight Neural Networks: Transforming Chip Design for Edge Device Deployment

Chaowei Gao

School of Integrated Circuit Science And Engineering,University Of Electronic Science And Technology of China, Chengdu, China
794127237@qq.com

Abstract. The continuous development and expanding applications of artificial intelligence are increasingly driving the demand for high computing power. Therefore, this article focuses on the impact and prospects of neural network models in chip design. This article mainly uses the research method of literature analysis to first explain the limitations of traditional chip architecture and discuss the impact of neural networks on chip design. Its impact is twofold: firstly, the development of neural networks has led to specialization in chip design; secondly, the auxiliary role played by neural networks in the process of chip design. This paper explores the advantages and potential of designing chips specifically for deploying lightweight neural network models. Furthermore, it proposes directions for the future development of neural networks and chip design. Lightweight neural network models have the advantages of light parameters, low latency, flexible deployment, and low hardware costs. They have significant advantages in the deployment of edge devices. Thanks to the government's policy support for technologies such as the Internet of Things and autonomous driving, a series of chips specifically designed for edge devices will emerge in the future, which will also promote the development of specialized chip design.

Keywords: Neural network, Chip design, Lightweight, artificial intelligence

1. Introduction

In recent years, artificial intelligence has developed rapidly, and its applications have penetrated into various industries, affecting many industries and will continue to do so. As an important component of artificial intelligence, neural networks have already formed a scale after decades of development since Geoffrey Hinton broke through the problem of gradient vanishing in traditional backpropagation algorithms in deep networks in the 1980s[1]. Its current applications are diverse, encompassing but not limited to facial recognition, intelligent driving, natural language processing, financial services, and other domains. Afterwards, a series of neural network models emerged successively, such as AlexNet and ResNet. And some models with specific application directions, such as lightweight MobileNet, EfficientNet, and ShuffleNet for edge devices. In this era of interdisciplinary integration, integrating chip design with neural networks has become one of the future research directions, and the two are mutually reinforcing. The synergy between neural networks and chip design is revolutionizing the field of computing. This study explores the evolving

DOI: 10.54254/2755-2721/2025.LD24520

relationship between these two domains, highlighting the limitations of conventional chip architectures in supporting the growing demands of neural network applications. It further examines the current integration of neural networks within the chip design process, with a particular focus on the burgeoning role of lightweight neural networks in edge computing. Finally, this research delineates the future trajectory of this dynamic field, aiming to foster the widespread adoption of lightweight neural networks in chip design and analyze their promising application prospects.

2. The limitations of traditional chip architecture in neural network applications

Neural network deployment platforms can be broadly categorized into general-purpose computing platforms and application-specific integrated circuit (ASIC) chips. General-purpose computing platforms encompass central processing units (CPUs), graphics processing units (GPUs), and field-programmable gate arrays (FPGAs), while ASICs primarily comprise digital and mixed-signal variants.

Platforms based on traditional von Neumann architecture, represented by CPUs and GPUs, suffer from severe memory performance lag and processor computing speed issues when deploying neural networks. Deploying on such platforms will waste a significant amount of computing power. In addition, general-purpose computing platforms also have the problem of high energy consumption. They are not a platform specifically designed for deploying neural networks, and there are insufficient issues in terms of running speed, energy consumption, and other aspects. FPGA-based acceleration platforms offer more flexible hardware resource scheduling solutions; however, the highly redundant circuit designs required to support their reconfigurable characteristics limit further optimization of their energy efficiency and speed performance [2].

In addition, as Moore's Law gradually deviates, the development of technology has encountered bottlenecks, limited by physical laws, and costs are increasing. So a better solution is to design chips specifically for deploying neural networks.

3. The influence of neural networks on chip design

In this era where the wave of artificial intelligence is sweeping the globe, chip design will inevitably undergo transformative changes due to artificial intelligence technology. The process of chip design will be assisted by neural networks. Secondly, the chip will be forced to be restructured in its architecture. Chip design is usually divided into six stages: functional design, architecture design, logic design, circuit design, physical design, and verification testing. In architectural design, neural networks can optimize the architecture, finding a balance between performance, power consumption, and structure. During logical design, neural networks can assist in completing a portion of the logic design. In physical design, neural networks can optimize wiring schemes and enhance efficiency.

3.1. The influence of current neural networks in chip design

The current influence of neural networks in chip design can be divided into two aspects: the development of neural networks promotes the specialization of chip design, and the auxiliary role played by neural networks in the chip design process affects the changes in chip design.

The specialization of chip design mainly focuses on hardware optimization and the search for new computing architectures. Regarding hardware, the chips utilized for AI deployment are predominantly ASICs and FPGAs. ASICs are known for their high performance and low power

DOI: 10.54254/2755-2721/2025.LD24520

consumption, making them particularly well-suited for neural network deployment and offering significant advantages in this domain. However, he also has the disadvantages of high development costs, long cycles, and lack of flexibility. FPGA has high flexibility, programmability, low latency, and parallel computing capability, which can be used in various scenarios. But its circuit is highly redundant, with high power consumption and low resource utilization. In addition to traditional electronic chips, using photonic chips as a neural network deployment platform is also a research direction. Photon chips, also known as optoelectronic chips or photonic integrated circuits, use light waves (electromagnetic waves) as carriers for information transmission or data processing. Photon chips generally employ optical waveguide media for signal transmission, integrating the modulation, transmission, demodulation, and processing of optical and electrical signals[3]. Their key characteristics include low power consumption, low latency, high efficiency, and the potential for miniaturization.

Neural networks can play an auxiliary role in the chip design process. In contemporary chip design, EDA tools are extensively utilized for layout optimization, while the generation of global architectures is another key aspect of the process. EDA layout can be assisted by reinforcement learning algorithms, greatly improving the efficiency of layout and routing, shortening the chip design cycle [4], and thus reducing costs.

3.2. The impact of lightweight neural networks on chip design

Chips on edge devices, including smartphones, drones, sensors, and other devices, have the characteristics of small storage space, low computing power, and real-time application requirements. Therefore, neural networks deployed on edge devices need to have lightweight characteristics.

The lightweight model also has certain limitations. If some methods only achieve ideal compression and acceleration in theory, they may not achieve the best results in practical deployment, which involves hardware support and compatibility issues [5]. But this practical limitation provides research value for lighter chip designs.

The current lightweight neural network models mainly include MobileNet, ShuffleNet, EfficientNet, and other models. The core principle of MobileNet is depthwise separable convolution. Depthwise separable convolution refers to the use of a large number of separable depth convolution operations to reduce the number of parameters. MobileNet [6] employs two hyperparameters, the width parameter and the resolution multiplier, to control the number of channels per layer and the image resolution, respectively. This approach further contributes to the network's lightweight nature. ShuffleNet [7] uses group convolution to reduce computational complexity and solves the problem of information blockage caused by group convolution through channel rearrangement. EfficientNet [8] adopts a composite scaling strategy, using mixed coefficients to uniformly scale width, depth, and resolution. The mixing coefficient is determined by neural architecture search (NAS), which specifies how resources are allocated to the width, depth, and resolution of the network to achieve a balance between lightweight and accuracy.

In addition to using lightweight neural network models, quantifying and pruning the original model is also a common method for compressing the model. At present, research on quantitative neural networks has received widespread attention. Quantification, as shown in Figure 1, is one of the commonly used methods in digital logic and has entered the field of neural networks. Its main goal is to compress neural network models without significantly reducing their performance. Quantization is the process of dividing continuous numbers distributed in the real field at certain intervals, replacing all values within a certain interval with a certain value. Pruning, as shown in Figure 2, involves removing some unimportant nodes to reduce computational or search complexity.

DOI: 10.54254/2755-2721/2025.LD24520

Pruning has great applications in many algorithms, such as decision trees, neural networks, search algorithms, database design, etc. Pruning effectively alleviates overfitting and reduces computational complexity in decision trees and neural networks, while also narrowing the search scope and improving efficiency in search algorithms.

By combining lightweight neural network models with pruning and quantization methods, the data processing load of the model is greatly reduced, meeting the fast response and high accuracy requirements of this type of equipment.

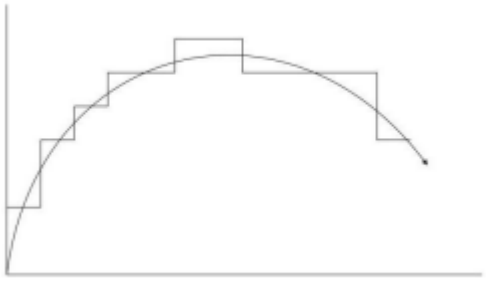

Figure 1: Quantization of continuous curves into discrete form

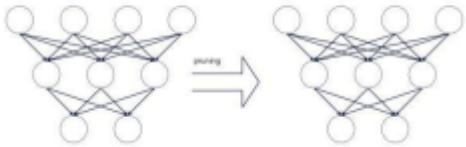

Figure 2: Pruning of unimportant parts from the network

4. Future development

As an upstream industry of the chip industry, chip design is crucial for the smooth and upward development of the entire chip industry. Now it seems that with the deployment of more and more neural networks, lightweighting neural networks to reduce their hardware costs has become a research hotspot, and correspondingly, solving problems from a hardware perspective. Note that numerous marginalized devices are also future development hotspots, such as the Internet of Things that has emerged with the popularization of 5G technology. The sensors required for the Internet of Things can also be organically combined with neural networks, making the IoT system more complete and reliable. The development of the low-altitude economy is contingent upon robust hardware support, while its real-time operational requirements necessitate the integration of neural networks. Based on the current reality, hardware resources are relatively limited, and for cost considerations, the deployment of lightweight neural networks will inevitably become a major direction for future research. This presents a practical demand for chips that can adapt to lightweight neural networks. Photon chips have great potential for performance improvement, as they do not require the pursuit of process limits like electronic chips, and therefore have the potential to serve as devices that continue Moore's Law. At present, photonic chips have frequently appeared in the development and investment plans of artificial intelligence in major countries around the world [9].

DOI: 10.54254/2755-2721/2025.LD24520

5. Conclusion

This study explores the impact and future prospects of neural networks on chip design. It analyzes the significant memory performance lag, inefficient utilization of computing resources, and high energy consumption associated with traditional chip architectures when deploying neural networks, thereby demonstrating their limitations. The influence of neural networks on chip design manifests in two primary aspects. Firstly, the evolution of neural networks drives the specialization of chip design, encompassing hardware optimization and the exploration of novel computing architectures. Secondly, neural networks play an auxiliary role in the chip design process, influencing design methodologies, such as employing neural network models for global architecture generation or EDA layout optimization. Furthermore, the article examines the current impact of neural networks on chip design and the specific influence of lightweight neural networks. Finally, considering current policy support, such as [mention specific policies if possible], and future development needs, the article discusses the future trajectory of chip design. It explores the application advantages and potential of lightweight neural network models in edge devices, highlighting the opportunities they present for advancements in chip design.

References

[1] Hinton, G. E. , Osindero, S. , & Teh, Y. -W. (2006). A Fast Learning Algorithm for Deep Belief Nets. Neural Computation, 1527–1554. https://doi. org/10. 1162/neco. 2006. 18. 7. 1527

[2] Liu W. (2024). Research on Hybrid Digital-Analog Deep Neural Network Systems. https://link. cnki. net/doi/10. 27005/d. cnki. gdzku. 2024. 005775doi:10. 27005/d. cnki. gdzku. 2024. 005775.

[3] SHEN Xiang, YU Jieping, WANG Li. Research Focus and Trend in Photonic Chip[J]. Frontiers of Data & Computing, 2023, 5(4): 3-15.

[4] Li yingyue. (2024). EDA Floorplan Algorithm Research Based on Deep Reinforcement Learning. https://link. cnki. net/doi/10. 26969/d. cnki. gbydu. 2024. 002629doi:10. 26969/d. cnki. gbydu. 2024. 002629.

[5] Huo Z, Zheng Y & Chen Y. (2023). Research progress on model lightweighting and acceleration, (03), 35-40.

[6] Howard, A. , Zhu, M. , Chen, B. , Kalenichenko, D. , Wang, W. , Weyand, T. , … Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv: Computer Vision and Pattern Recognition, arXiv: Computer Vision and Pattern Recognition.

[7] Zhang, X. , Zhou, X. , Lin, M. , & Sun, J. (2018). ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT. https://doi. org/10. 1109/cvpr. 2018. 00716

[8] Tan, M. , & Le, QuocV. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.

[9] Zhang Yi, Zhu Xiaoling, Yuan Liu, Guo Xiaolei & Yan Hongyan. (2025). Analysis of the Main Developments of Artificial Intelligence Chip Technology in 2024. Unmanned systems technology, 8 (01), 108-116. The doi: 10. 19942 / j. i SSN. 2096-5915. 2025. 01. 09.