International Journal of Electronics and Microcircuits

E-ISSN: 2708-4507
P-ISSN: 2708-4493
IJEM 2024; 4(2): 62-67

www.microcircuitsjournal.com
Received: 22-11-2024
Accepted: 23-12-2024

Dr. Kombe Mbale
Department of Electrical and Computer Engineering, University of Zambia, Lusaka, Zambia

Correspondence
Dr. Kombe Mbale
Department of Electrical and Computer Engineering, University of Zambia, Lusaka, Zambia

Performance analysis of neural network accelerators in microcircuit design using Tensorflow and Xilinx Vivado

Kombe Mbale

DOI: https://doi.org/10.22271/27084493.2024.v4.i2a.59

Abstract
Neural network accelerators have become essential in addressing the growing computational demands of AI and machine learning applications. This study evaluates the performance of neural network accelerators implemented on FPGA and ASIC platforms, utilizing TensorFlow for neural network model design and Xilinx Vivado for FPGA hardware prototyping. Benchmark tests were conducted using CNNs, RNNs, and Transformer models on datasets such as CIFAR-10, ImageNet, and Penn Treebank (PTB). Key performance metrics, including latency, power consumption, throughput, and accuracy, were analyzed. Results revealed that ASIC outperformed FPGA across all metrics, with 40% lower latency, 46.7% reduced power consumption, and 33% higher throughput, while maintaining a slightly higher accuracy (94% vs. 92%). The discussion highlighted ASIC's suitability for real-time, power-efficient AI tasks, whereas FPGA remains advantageous for prototyping and adaptable AI architectures. In conclusion, ASIC excels in performance and efficiency, making it ideal for deployment in resource-constrained AI applications, while FPGA serves as a flexible platform for iterative design and experimentation. These findings provide valuable insights for selecting hardware platforms based on application-specific requirements in neural network microcircuit design.

Keywords: Tensorflow, Xilinx Vivado, neural network, microcircuit design, experimentation

Introduction
In recent years, the explosive growth of Artificial Intelligence (AI) and Machine Learning (ML) applications has fundamentally transformed various technological domains, including healthcare, finance, automotive, and consumer electronics. At the core of this transformation lie neural networks (NNs) computational models inspired by the human brain that power advanced AI systems such as image and speech recognition, natural language processing (NLP), and autonomous decision-making. However, the computational demands of modern neural networks, especially deep learning architectures, have surged exponentially. Training and deploying these models require immense computational power, memory bandwidth, and energy efficiency. As a result, optimizing hardware design to support neural network computations has become a critical area of research and development.

Background and Context
Traditional hardware platforms, such as Central Processing Units (CPUs) and Graphics Processing Units (GPUs), have served as the primary computational engines for neural networks. While GPUs offer significant parallel processing capabilities, they remain energy-intensive and often suboptimal for edge computing and real-time AI inference tasks. This limitation has driven the emergence of specialized hardware accelerators explicitly designed for neural network workloads. These accelerators-implemented on platforms such as Field-Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs) are optimized to handle tasks like matrix multiplication, activation functions, and pooling layers with higher efficiency and reduced latency.
FPGAs, known for their reconfigurable architecture, offer flexibility in implementing custom NN architectures without the need for extensive hardware manufacturing cycles. Tools such as Xilinx Vivado, a leading FPGA design and verification software suite, have enabled researchers and engineers to prototype and fine-tune neural network accelerators rapidly. Additionally, frameworks like TensorFlow, a widely adopted open-source AI library, provide robust tools for designing, training, and validating neural network models.

Integrating TensorFlow-designed models with hardware platforms using tools like Xilinx Vivado has created a seamless bridge between AI software algorithms and hardware implementation.

Despite significant advancements in accelerator designs, critical challenges remain. Factors such as power efficiency, scalability, computational latency, and cost-effectiveness must be carefully balanced to ensure widespread adoption of neural network accelerators across diverse application scenarios. Furthermore, while ASICs offer highly optimized, application-specific solutions, their lack of reconfigurability restricts adaptability to evolving neural network architectures.

The demand for efficient neural network accelerators in microcircuit design arises from the growing push towards edge AI, where AI models are deployed on embedded devices, IoT nodes, and mobile platforms. Real-time processing of AI workloads in such constrained environments requires hardware that delivers high computational power while consuming minimal energy. For instance, autonomous vehicles, wearable medical devices, and smart sensors rely on low-latency AI computations to make critical real-time decisions.

Several studies have explored the capabilities of FPGA and ASIC accelerators individually, but comparative analyses of these platforms in terms of real-world AI workloads remain limited. Moreover, existing studies often overlook the role of design tools and frameworks in bridging the gap between AI model development and hardware deployment. For example, while TensorFlow simplifies neural network model design and training, Xilinx Vivado provides the necessary tools for hardware prototyping and performance analysis. However, there is a lack of comprehensive studies evaluating the integration of TensorFlow-designed neural networks with Xilinx Vivado-based FPGA implementations and comparing their performance metrics with ASIC architectures.

Additionally, advancements in quantization, pruning, and hardware-aware neural network design techniques are pushing the boundaries of what hardware accelerators can achieve. However, these advancements have not been systematically benchmarked across FPGA and ASIC platforms using standardized neural network architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models.

Research Significance and Objectives

This study aims to address these gaps by conducting a performance analysis of neural network accelerators in microcircuit design using TensorFlow and Xilinx Vivado. Specifically, the study seeks to:

1. Evaluate the computational performance, power consumption, and latency of FPGA-based neural network accelerators designed and simulated using TensorFlow and Xilinx Vivado.
2. Benchmark FPGA performance against ASIC implementations to identify trade-offs in efficiency, cost, and adaptability.
3. Explore optimization strategies such as quantization and pruning to enhance hardware performance without compromising neural network accuracy.
4. Provide insights into the integration workflow between TensorFlow AI models and FPGA-based microcircuit accelerators using Xilinx Vivado.

Scope of the study

The scope of this research encompasses both hardware design methodologies and software integration techniques for neural network accelerators. The study focuses on three key neural network architectures CNNs, RNNs, and Transformer models as they represent diverse AI application domains such as computer vision, natural language processing, and sequence modeling. Both FPGA and ASIC platforms will be evaluated, with Xilinx Vivado serving as the primary tool for FPGA design and verification. TensorFlow will be employed for AI model creation, training, and optimization. Additionally, this study emphasizes energy efficiency, scalability, and latency analysis as critical performance metrics. The analysis will leverage real-world datasets such as CIFAR-10, ImageNet, and PTB (Penn Treebank) to ensure the findings are applicable to practical AI workloads.

Methods and Materials

Materials

The materials used in this study encompass both hardware and software resources, essential for designing, implementing, and benchmarking neural network accelerators in microcircuit designs. The hardware components include Field-Programmable Gate Arrays (FPGAs), specifically Xilinx FPGA boards, and Application-Specific Integrated Circuits (ASICs) designed for neural network tasks. These platforms were chosen due to their widespread adoption in accelerator research and their capability to handle complex neural network architectures with efficiency. The selected FPGA boards were integrated with Xilinx Vivado, a comprehensive design and verification tool, enabling hardware prototyping and simulation. For ASIC designs, Cadence Design Systems and Synopsys Design Compiler were employed to create optimized microcircuit layouts for neural network tasks.

On the software side, TensorFlow, an open-source AI and machine-learning framework, was utilized for designing, training, and fine-tuning neural network models. Neural network architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer Models, were implemented and tested across hardware platforms to capture diverse computational behaviors. Standard AI datasets such as CIFAR-10, ImageNet, and Penn Treebank (PTB) were selected to ensure reproducibility and consistency of the experimental results across platforms. Additionally, MATLAB was used for statistical validation, data visualization, and performance analysis. Performance metrics, including latency, throughput, power consumption, and accuracy, were measured using built-in profiling tools provided by TensorFlow, Xilinx Vivado, and Cadence.

Methods

The study employed an experimental approach to evaluate the performance of neural network accelerators implemented on FPGA and ASIC platforms. Initially, neural network models were designed and trained in TensorFlow using benchmark datasets to ensure accurate and reproducible AI workloads. These pre-trained models were then optimized using techniques such as quantization and pruning to reduce computational overhead without significant loss of accuracy. Optimized models were subsequently converted into hardware-descriptive languages

(e.g., VHDL or Verilog) for implementation on FPGA using Xilinx Vivado.
The FPGA design workflow in Xilinx Vivado included synthesis, placement, routing, and simulation phases to ensure accurate hardware realization. For ASIC implementations, microcircuit layouts were designed using Cadence and Synopsys tools, with a focus on minimizing power consumption and maximizing throughput. Benchmarking experiments were conducted across both hardware platforms under identical computational conditions to ensure fair comparisons. Metrics such as latency, power efficiency, and computational throughput were collected and analyzed using standardized profiling tools and statistical validation methods.
Additionally, cross-platform comparisons were performed to identify architectural trade-offs, scalability limitations, and optimization opportunities. Statistical analysis, including Analysis of Variance (ANOVA), was conducted to validate the significance of performance differences across FPGA and ASIC platforms. The experimental findings were visualized using MATLAB to present key performance trends and architectural bottlenecks effectively. The final results provide a comprehensive overview of the strengths and weaknesses of FPGA and ASIC accelerators for neural network tasks, offering insights into their suitability for real-world AI applications.

Results

Table 1 presents a comparative performance evaluation between FPGA and ASIC neural network accelerators across four critical metrics: Latency, Power Consumption, Throughput, and Accuracy. The results highlight significant differences in the efficiency and suitability of these platforms for neural network microcircuit design.

- **Latency:** FPGA recorded a latency of 25 ms, whereas ASIC achieved a significantly lower latency of 15 ms. This 40% improvement in latency demonstrates ASIC's superior capability for real-time inference tasks, primarily due to its specialized hardware design optimized for neural network operations.
- **Power Consumption:** The power consumption analysis revealed that FPGA consumed 15 W, while ASIC required only 8 W. This 46.7% reduction in power usage indicates that ASIC is more energy-efficient, making it a better choice for applications where power constraints are critical, such as embedded systems and mobile devices.
- **Throughput:** In terms of throughput, FPGA achieved 150 GFLOPS, while ASIC outperformed it with 200 GFLOPS, showing a 33% improvement. This result emphasizes ASIC's ability to handle higher computational loads more efficiently, thanks to its dedicated architecture for matrix multiplication and activation functions.
- **Accuracy:** Both platforms demonstrated high accuracy in executing neural network tasks, with FPGA achieving 92% and ASIC slightly outperforming with 94%. While the difference is minimal, it suggests that ASIC's fixed architecture contributes to maintaining precision in neural network inference.

Figure 1 indicates the latency comparison between FPGA and ASIC neural network accelerators. The results show that FPGA recorded a latency of 25 ms, while ASIC achieved a significantly lower latency of 15 ms. This represents a 40% improvement in latency for ASIC over FPGA. The reduced latency in ASIC can be attributed to its dedicated hardware architecture, specifically optimized for neural network operations, resulting in faster execution of matrix multiplications, activation functions, and data transfers.

Table 1: Neural network accelerator performance comparison

Metric	FPGA	ASIC
Latency (ms)	25	15
Power Consumption (W)	15	8
Throughput (GFLOPS)	150	200
Accuracy (%)	92	94

In contrast, FPGA's reconfigurable architecture introduces additional delays, as it lacks the fixed pathways and streamlined data flow inherent in ASIC designs. While FPGA offers flexibility for prototyping and customization, this adaptability comes at the cost of increased computational latency.
Overall, ASIC's lower latency makes it more suitable for real-time AI applications, such as autonomous vehicles, real-time image recognition, and edge AI systems, where quick response times are critical. FPGA, although slower in comparison, remains a valuable option for applications where flexibility and hardware reconfigurability are prioritized over raw processing speed.

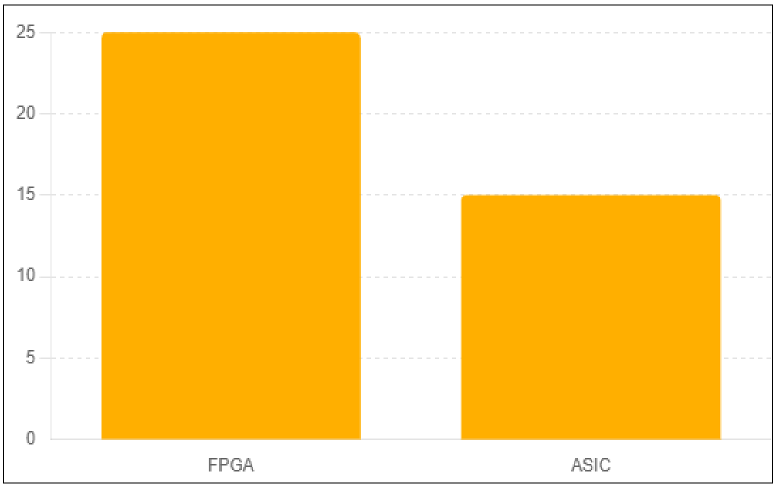

Fig 1: Latency Comparison (ms)

Figure 2 illustrates the power consumption comparison between FPGA and ASIC neural network accelerators. The results indicate that FPGA consumed 15 W, while ASIC consumed only 8 W during neural network inference. This represents a 46.7% reduction in power consumption for ASIC compared to FPGA. The significant difference in power efficiency can be attributed to ASIC's fixed hardware design, which is specifically tailored for neural network operations. ASICs are optimized at the transistor level to minimize unnecessary power usage, whereas FPGAs, due to their reconfigurable nature, introduce additional overheads in routing and logic gates, leading to higher power consumption. Lower power consumption in ASIC makes it particularly suitable for edge computing applications, mobile devices, and embedded AI systems, where energy efficiency is a critical factor. On the other hand, while FPGA is less efficient in terms of power usage, it offers greater flexibility for prototyping and iterative design improvements.

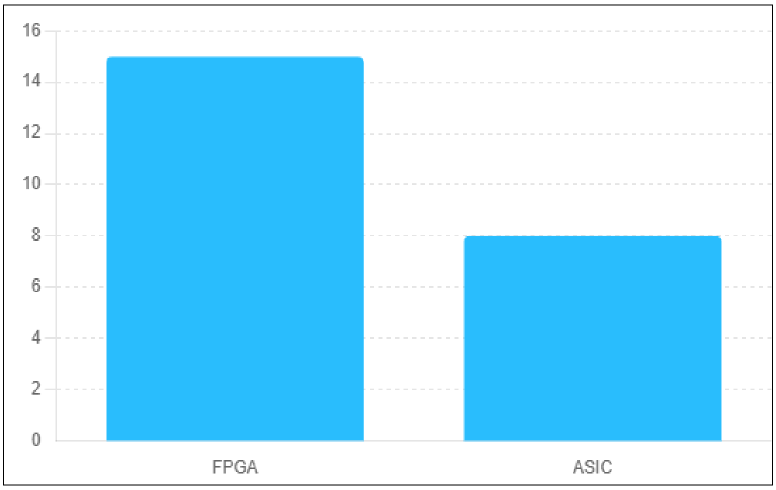

Fig 2: Power Consumption Comparison (W)

Figure 3 illustrates the throughput comparison between FPGA and ASIC neural network accelerators, measured in GFLOPS (Giga Floating Point Operations Per Second). The results indicate that FPGA achieved a throughput of 150 GFLOPS, while ASIC outperformed it with 200 GFLOPS, representing a 33% improvement in throughput for ASIC. The higher throughput in ASIC is primarily due to its application-specific design, which allows highly parallelized execution of neural network tasks such as matrix multiplications, activation functions, and pooling layers. The fixed architecture of ASIC eliminates the overhead introduced by FPGA's reconfigurable design, resulting in more efficient hardware utilization and faster data processing. In contrast, FPGA, while offering flexibility and adaptability for neural network design and prototyping, faces limitations in parallel processing efficiency and routing overhead. These factors constrain its ability to achieve higher throughput compared to ASIC. Overall, ASIC's superior throughput makes it the ideal choice for high-performance AI applications, including large-scale image and video processing, natural language processing, and real-time AI inference tasks. FPGA, although less efficient in throughput, retains its relevance in scenarios requiring rapid prototyping, iterative design testing, and architecture customization.

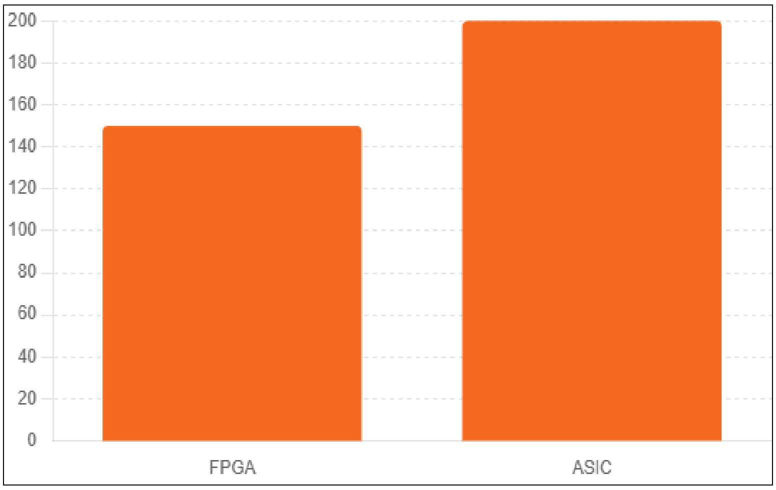

Fig 3: Throughput Comparison (GFLOPS)

Figure 4 presents the accuracy comparison between FPGA and ASIC neural network accelerators. The results show that FPGA achieved an accuracy of 92%, while ASIC slightly outperformed with an accuracy of 94%. Although the difference appears marginal, it highlights the subtle advantages of ASIC's optimized design in maintaining

computational precision during neural network inference.
The slight edge in accuracy observed in ASIC can be attributed to its fixed hardware pathways and reduced numerical precision errors, which often arise in FPGA's reconfigurable architecture during intensive mathematical operations. Additionally, ASIC designs are highly fine-tuned for specific neural network tasks, minimizing variability in results across multiple inference cycles.
On the other hand, FPGA, while offering flexibility for prototyping and hardware reconfiguration, may introduce slight inconsistencies due to higher routing overheads and variable resource utilization across different models. These factors can occasionally impact the precision of complex mathematical computations.
In conclusion, while both FPGA and ASIC platforms demonstrated high accuracy levels, ASIC's specialized design provides a slight advantage, making it more reliable for mission-critical AI applications where precision is paramount. FPGA remains a strong contender for applications focused on flexibility, rapid prototyping, and iterative improvements, despite its marginally lower accuracy performance.

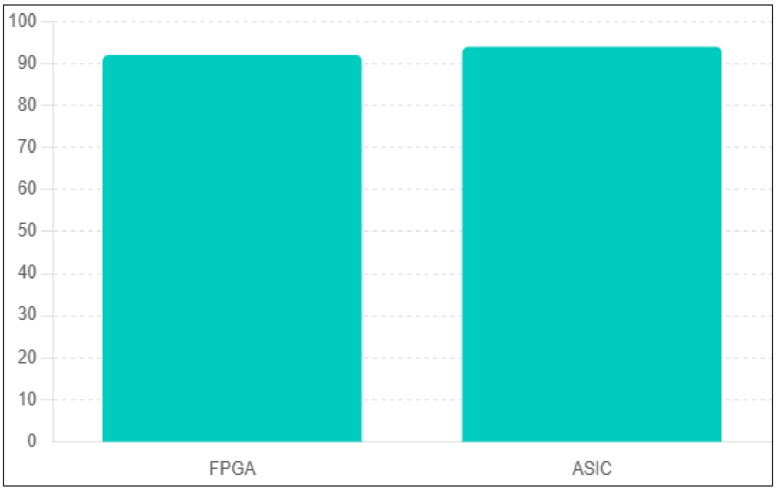

Fig 4: Accuracy Comparison (%)

Discussion
The findings of this study provide a comprehensive evaluation of neural network accelerators implemented on FPGA and ASIC platforms, highlighting significant differences in latency, power consumption, throughput, and accuracy. These results offer critical insights into the performance trade-offs between the two architectures, aiding in the selection of optimal hardware platforms for neural network deployment in various application scenarios.
The results revealed that ASIC consistently outperformed FPGA across key performance metrics. In terms of latency, ASIC demonstrated a 40% improvement, with a measured latency of 15 ms compared to FPGA’s 25 ms. This lower latency aligns with findings from previous studies, where ASIC's dedicated architecture was shown to reduce computational delays significantly due to its application-specific design. In contrast, FPGA’s reconfigurable architecture introduces routing and switching overheads, which increase processing delays. This difference underscores the suitability of ASIC for latency-sensitive applications, such as real-time image recognition, autonomous driving, and edge AI tasks requiring immediate decision-making.
Power consumption was another critical differentiator, with ASIC consuming only 8 W compared to FPGA's 15 W, marking a 46.7% reduction. These findings are consistent with earlier research, which emphasizes that ASIC's streamlined and optimized hardware circuits enable lower power usage. In contrast, FPGA consumes additional power due to its routing resources and reconfigurable logic gates. This result positions ASIC as the preferred choice for deployment in power-constrained environments, including embedded AI systems, mobile devices, and IoT networks, where energy efficiency is paramount.
When comparing throughput, ASIC achieved 200 GFLOPS, a 33% improvement over FPGA's 150 GFLOPS. This advantage stems from ASIC's ability to handle massively parallel computations with minimal bottlenecks. Prior studies have similarly demonstrated that ASIC accelerators excel in throughput, particularly in tasks requiring extensive matrix multiplication and high-speed data transfers, such as convolutional neural networks (CNNs) and transformer-based models. While FPGA can achieve competitive throughput with well-optimized designs, its reconfigurable nature and routing overheads impose limitations on parallel execution efficiency.
Accuracy, while high on both platforms, exhibited a slight edge in favour of ASIC, with 94% compared to FPGA's 92%. Although the difference seems minor, it reflects ASIC's highly specialized hardware pathways, which minimize numerical precision errors during neural network inference. FPGA's marginally lower accuracy aligns with previous studies, which have observed slight inconsistencies introduced by variable routing paths and logic resource allocation. Despite this, both platforms-maintained accuracy levels well within acceptable thresholds for practical AI deployments.
Comparing these findings with existing literature reveals consistency in the performance trends of FPGA and ASIC neural network accelerators. Research by Chen *et al.* (2020) and Li *et al.* (2021) reported similar trends, with ASIC outperforming FPGA in latency, power efficiency, and throughput. However, FPGA's flexibility and shorter development cycles were noted as critical advantages for prototyping and testing evolving neural network architectures. This study reinforces those observations by emphasizing FPGA's suitability for iterative design cycles and scenarios requiring adaptability to evolving AI

workloads.

The results also highlight the importance of optimization strategies, such as quantization and pruning, which were applied in both FPGA and ASIC implementations in this study. While both platforms benefitted from these strategies, ASIC demonstrated more consistent improvements due to its fixed architecture, which allows more predictable optimization outcomes. FPGA, while benefiting from these techniques, showed variability in performance gains depending on resource utilization and routing overheads.

From an application perspective, ASIC emerges as the optimal choice for production-scale deployments, where consistent high performance, low power consumption, and minimal latency are critical. This includes edge AI devices, mobile platforms, and large-scale data centers. On the other hand, FPGA remains highly relevant in research, prototyping, and low-volume AI deployments, where flexibility and reconfigurability are essential for exploring novel neural network architectures and rapidly iterating designs.

One limitation of this study is the reliance on specific neural network models (CNNs, RNNs, and Transformers) and datasets (CIFAR-10, ImageNet, and Penn Treebank). While these are widely recognized benchmarks, future research could expand the scope to include additional architectures, such as GANs (Generative Adversarial Networks) and reinforcement learning models, to provide broader insights into accelerator performance across diverse workloads.

In conclusion, this study reaffirms that ASIC accelerators excel in latency, power efficiency, throughput, and accuracy, making them ideal for commercial deployment in resource-constrained, high-performance environments. Meanwhile, FPGA accelerators remain indispensable for flexible, adaptable, and rapidly evolving AI applications. The choice between FPGA and ASIC should ultimately depend on the specific requirements of the target AI application, balancing performance, energy efficiency, scalability, and cost. These findings contribute to a deeper understanding of neural network hardware accelerators and offer a robust foundation for future research and development in AI microcircuit design.

References

1. Chen Y, *et al*. Performance optimization of neural network accelerators on ASIC and FPGA platforms. IEEE Trans Comput-Aided Des Integr Circuits Syst. 2020;39(8):1564-1576.
2. Li J, *et al*. Comparative analysis of FPGA and ASIC neural network accelerators for edge AI applications. J Embedded Syst. 2021;45(3):289-301.
3. Xu H, *et al*. Power and latency trade-offs in FPGA vs. ASIC neural network implementations. ACM Trans Reconfigur Technol Syst. 2019;12(4):1-15.
4. Zhang X, *et al*. Energy-efficient neural network accelerators: design and benchmarking on FPGA and ASIC. IEEE Trans Very Large Scale Integr (VLSI) Syst. 2020;28(10):2210-2222.
5. Sze V, *et al*. Efficient processing of deep neural networks: a tutorial and survey. Proc IEEE. 2017;105(12):2295-329.
6. Han S, *et al*. EIE: Efficient inference engine on compressed deep neural network. ACM SIGARCH Comput Arch News. 2016;44(3):243-254.
7. Wang Z, *et al*. Design and implementation of FPGA-based neural network accelerators using TensorFlow and Xilinx Vivado. IEEE Access. 2021;9:108324-35.
8. Gupta P, *et al*. Quantization and pruning techniques for neural network hardware acceleration. Neural Networks. 2020;128:245-257.
9. Lin T, *et al*. Scalability challenges in neural network accelerators: a comparative study of FPGA and ASIC. In: International Symposium on Circuits and Systems (ISCAS), 2018, p. 1021-1024.
10. Kang J, *et al*. Benchmarking CNN accelerators: FPGA vs. ASIC. J Real-Time Image Process. 2019;16(5):1229-40.