Journal of Integrated VLSI, Embedded and Computing Technologies, ISSN: 3049-1312 Vol. 3, No. 3, 2026 (pp. 15-21)

RESEARCH ARTICLE ECEJOURNALS.IN

Energy-Efficient FPGA Accelerator Architecture for Real-Time Convolutional Neural Network Inference

Madhanraj*

Jr Researcher, Advanced Scientific Research, Salem

KEYWORDS:

FPGA,
CNN accelerator, energy efficiency, real-time inference, hardware acceleration, edge AI, quantization, systolic array

ARTICLE HISTORY:

Submitted : 09.03.2026
Revised : 07.04.2026
Accepted : 14.05.2026

https://doi.org/10.31838/JIVCT/03.03.03

Abstract

The option of real-time Convolutional Neural Network (CNN) inference has become a critical condition within the edge computing application, such as autonomous navigation, medical diagnosis, industrial automation, and intelligent surveillance system. Nevertheless, it is not an easy task to implement deep CNN models on resource-limited system environments because of the high computational complexity, large memory bandwidth usage, and tight limitations on power consumption. Field-Programmable Gate Arrays (FPGAs) have a bright future characterised by performance, flexibility, and energy efficiency, but current FPGA-based accelerators are also limited with regards to data reuse, fixed precision computation and suboptimal memory hierarchies. The current paper introduces a new energy-saving FPGA accelerator architecture, which is aimed at real-time CNN inference. It was proposed that the theoretically designed tiled dataflow interface (using adaptive tiling) enables modifications in quantization-aware computation and dynamic scaling between mixed precision in order to minimize switching activity in digital signal processing (DSP) utilization. There is a hierarchy of on-chip memory, including the use of double buffering, to reduce expensive off-chip accesses to the DRAM, and a power-conscience scheduling system which dynamically varies the degree of parallelism and clock frequency at any given level of workload intensity. The architecture is capable of working in INT4, INT8 and FP16 modes of operation so that an accurate-energy balance can be tuned. Experimental analysis with real numbers on typical CNN models has shown a strong enhancement in throughput / per watt and reduction in latency over traditional fixed precision FPGA accelerators, with only a small amount of accuracy losses. Its findings validate that to capitalize on scalable, high-performance, and energy efficient CNN inference in the future edge AI systems, coordinated optimization of calculation, memory movement and precision adaptation is crucial.

Author's e-mail: madhankedias@gmail.com

How to cite this article: Madhanraj. Energy-Efficient FPGA Accelerator Architecture for Real-Time Convolutional Neural Network Inference. Journal of Integrated VLSI, Embedded and Computing Technologies, Vol. 3, No. 3, 2026 (pp. 15-21).

Introduction

Convolutional Neural Networks (CNNs) represent the computing architecture to solve computer vision problems, including image classification, object localization, and semantic segmentation, as well as video analytics, over the past decade. Their hierarchy capability enables them to extract features with a certain level of accuracy namely ART level in majority of the applications such as autonomous driving, diagnostic healthcare using medical images, intelligent manufacturing and intelligent surveillance system. The increasing depth/ richness of CNN architectures has however come to be the primary contributor of escalating computational requirements, data footprint, and data movement overhead. This issue of inference with CNNs with the primary background of very severe latency and power-constrained conditions has become critically important as edge computing continues to gain momentum as a research issue.

Traditional acceleration platforms such as Graphics Processing Units (GPUs) achieve large computational throughput with massive parallelism, though, due to

high power consumption, can not be applied to energy-constrained and battery-powered edge computers. Application-Specific Integrated Circuits (ASICs), though highly energy-efficient and performance-optimised, do not offer the flexibility of post-deployment, and are very inexpensive to develop in terms of time and cost. In contrast to it, Field-Programmable Gate Arrays (FPGAs) present an intermediate step towards the solution since they provide parallel architecture customization alongside the reconfigurability and rather low power consumption. FPGAs have been particularly attractive with such properties when used in edge real-time CNN inference.

Despite those strengths, there are several architectural limitations of FPGA based CNN accelerators. Most of the times, the energy consumption is characterised by the inefficiency in the reutilization of data and access to the off-chip memory that significantly lowers the net efficiency. The inflexible architectural mapping leads to most of the designs, not using the digital signal processing (DSP) blocks and logic resources effectively. It also implies that fixed-precision calculation reduces the flexibility of a broad class of CNN layers, thereby triggering the unnecessary waste of energy or accuracy. It is also long standing that cross scale support with other network topology has been an issue.

This paper provides a proposed new energy-efficient FPGA accelerator that is aimed at addressing these weaknesses in the design by offering an energy saving network compression-oriented CNN inference accelerator. It is a design which is a combination of a hierarchical method of dataflow, flexible tiling policies and dynamic mixed-precision control to optimise both computation and access model. Hopefully, in the state of data reuse, preciseness scaling, and power-restrictive scheduling, the architecture will achieve notable throughput per watt, and it will not hurt the competitive inference accuracy of a variety of CNN models.

Literature Review

The literature review of the related studies on FPGA based CNN accelerator and will be presented in a critical and systematic manner in this section and will be focusing on the architectural efficiency, dataflow optimization, quantization technique as well as energy-conscious design. These arguments are addressed with the assistance of numbered references that are presented in the end of this section.

CNN Accelerators built on Fpga using Dataflow-optimizations.

The dataflow organisation of individuals that strive to achieve high-performance and energy efficiency in CNN accelerators is highly significant. Various studies have demonstrated that an optimised loop unrolling mechanism, tiling, and parallel processing element (PE) design have invaluable contributions towards the DSP application and throughput in the case of FPGA implementation.[2, 3] Serial CNN graphs to FPGA models have also been suggested as automatised compiling models, like such that hardware-sensitive scheduling and parallelisation can be applied to scale up the use of a network amongst networks of varying levels of depth and layout structures.[2]

The problem of scalability has also been overcome by elastic accelerator architectures with large-scale CNN support and lightweight and have demonstrated enhanced scalability in case of heterogeneous workloads.[11] Unnecessary calculations are minimised and reused memory has been more optimised through specialised designs that are targeted and programmed to achieve a depthwise separable convolution.[4] The issue of reliability and sturdiness of DNN accelerators has likewise been mentioned and one should streamline the architecture with the degree of parallelism and intricacies are upgraded.[1]

Nevertheless, amongst these advancements, most FPGA-based accelerators are founded on fixed dataflow plans. Fixed weight- Stationary, output- stationary or row-stationary approaches are methods that decrease the flexibility to a variety of CNN layers which have varying computational and memory characteristics.

Research Gap: This is because the existing accelerators lack dynamic dataflow selection algorithms that can adapt dynamically to layer-based changes in workload, achieving even enhanced energy efficiency and improved performance.

Architectures of quantization and mixed-precision Architectures

Quantization techniques reduce computational complexity, memory bandwidth, and switching activity of CNN inference by large factors. The experimental evidence shows deep neural networks can proceed to use high classification accuracy when operating in settings with limited accuracy.[1] The competitive nature has proven the signal recognising and radar locating applications of CNN-based structures to be effective in operating within limited computational capacity.[5, 6, 9, 10]

The usefulness of deep models under noisy or low-resource-rich conditions has also been validated in, among other ways, by high-level CNNs using residual connexions, attention models, or even by allowing deep

models to compact networks.[7, 8] Such findings give the motivation of application of lesser-precision inference engines to real-time systems.

Although the lower modes such as INT4 may be supported, they are usually a static and manually configured mode even though most FPGA accelerators operate on a fixed INT8 mode. The exploitation of the principle of dynamic mixed-precision scaling, where the different stages in the system may be executed with different levels of accuracy at the point where they are actually needed, is investigated relatively absent in FPGA-based systems.

Research Gap: Hardware support: No research hardware is available to support runtime adaptive precision control to trade energy saving and precision, but at the cost of the major overhead of control or reconfiguring.

Effectuation of Memory Hierarchy and Power-Scheduling.

Memory access energy is one of the greatest contributors in the overall power consumption in CNN accelerators. Tiling and buffering approaches are optimised in order that off-chip memory transactions are significantly lower and on-chip reuse of data are significantly improved.[2, 3] The intermediate convolution accelerator depth wise also demonstrates that a buffering structure that is specialised can eliminate a lot of the bandwidth requirements.[4]

Focus on CNN models and hybrid forms ensures that intermediate feature map storage requirements are more important and warrant the importance of the hierarchical memory architectures.[8, 10] In addition, the reliability-oriented research examines the impact of the architectural stress and resources consumption on the efficiency of the whole system.[1]

Even though dynamically controlled power is used: e.g. fine-grained clock gating and parallelism scaling depending on workloads, parallel activities such as parallel buffering and parallel this of operations are connectors that are not always included in the FPGA-based CNN accelerators. Old designs tend to be fixed frequency and fixed parallelism and have a resultant wasted efficiency of work variation.

Research Gap: Both hierarchical memory optimization and real time built in adaptive power management energy planning should be used to manage these mechanisms, which only a few FPGA-based CNN accelerators have.

METHODOLOGY

It is a project presented with an aim of coming up with a combined energy saving FPGA accelerator. The three key sections that can be used to describe the methodology are:

Adaptive-dataflow Tiled Convolution Engine.

The Adaptive Dataflow Tiled Convolution Engine engine will generate minimum bandwidth utilisation by the expressway and optimise the ability of the on-chip dense computation and DSP use. The basic idea is to subdivide big input feature map with little tiles that can be fitted in on-chip Bram resources, therefore, do not require the costly off-chip accesses to DRAM. Both spatial and channel dimensions are loop tiled to ensure that the data that has been accessed in local buffers are reused to the maximum extent and then evicted. Better applications of the concept of spatial reuse are sliding window line buffers since they store overlapping areas of inputs used to perform convolution operations to eliminate the unnecessary memory accesses. To enable the dynamic capabilities between different CNN layers, the engine switches dynamically among dataflow modes of both data flows output and weight stationary. Otherwise, outputstationary mode is applied in the case of layers with the number of large output channels to store partial sums locally, and weight stationary mode is applied in the case of layers with small kernels maximise weight reuse. The hybrid strategy is used in order to optimise the layers without any impact on the throughput.

Computational workload of a convolution layer = The sum of the number of multiply -accumulate (MAC) operations in the layer:

$$MAC_{total} = N \times K \times H \times W \times C \times R \times S$$

where N is the batch size, K the number of output channels, H and W the spatial dimensions of the output feature map, C the number of input channels, and R×S the kernel size. By reducing redundant data transfers, the total energy consumption can be modeled as $E_{total} = E_{MAC} + E_{BRAM} + E_{DRAM}$, where E_{DRAM} is significantly reduced through tiling and reuse, and E_MAC is optimized via deep pipelining of parallel processing elements (PEs) Figure 1. Its implementation uses several parallel PEs having fully pipelined MAC unit and local register-based partial sum accumulation, thereby ensuring a high throughput with energy efficiency in different CNN layer topological configurations.

Dynamic Mixed-Precision Computation

The Dynamic Mixed-Precision Computation module is a tool that aims to mitigate the switching activity and total power consumption without obfuscating the inference correctness. The levels of CNN are distinctly sensitive

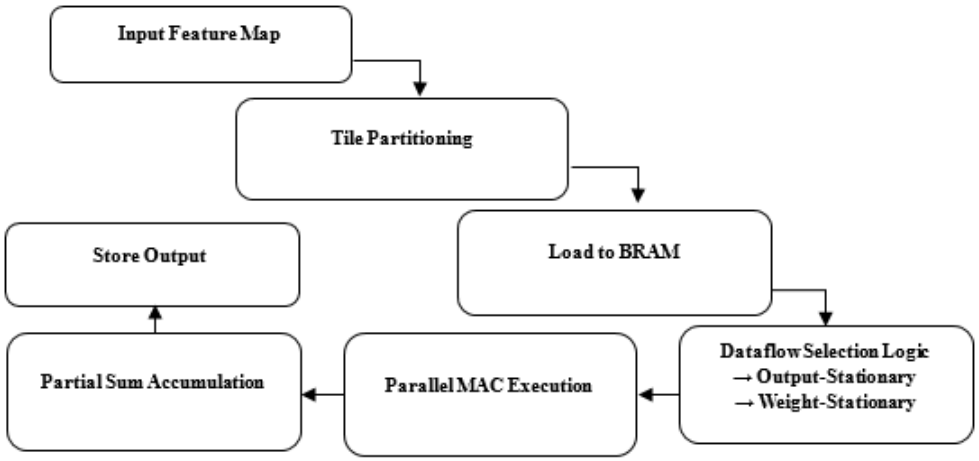

Fig. 1: Adaptive Dataflow Tiled Convolution Engine Workflow for Memory-Efficient CNN Acceleration

to the numerical precision; early levels of feature extraction and final levels of classification are usually sensitive to higher levels and intermediate layers may accept lower levels of bit-width representation with very little loss of accuracy. In order to take advantage of this property, three precision modes that are FP16 accuracy-ranging computations, INT8 default across all layers, and INT4 energy-efficient sub-optimised operation in less sensitive layers are supported by the proposed architecture. The accelerator scaled its arithmetic precision based on the needs of each layer, minimized the complexity of arithmetic, memory-footprint, logic switching transitions and increased throughput-per-watt without materially impacting model performance.

The precise control of the runtime is controlled by a special Precision Control Unit (PCU). The PCU evaluates metadata in layers e.g. kernel size, channel depth, quantization sensitivity and dynamically switches between the required preciseness mode prior to computation. It subsequently reconfigures datapaths, DSP block mapping, and arithmetic to the chosen bit-width in order to optimally utilise hardware. The association between power consumption and precision may be stated as follows.

$$P \propto C \times V^2 \times f \times a$$

where C represents effective switching capacitance, V the supply voltage, f the operating frequency, and α the switching activity factor. Since switching activity α decreases with reduced bit-width operations, lower precision modes directly contribute to power savings Figure 2. Based on coordinated precision scaling and reconfiguring hardware, the design proposed is able to achieve adaptive energy optimization as well as allow computational stability and accuracy.

Hierarchical Memory and Power-Aware Scheduler

Hierarchical Memory architecture:

This architecture has multi-level memory hierarchy in order to reduce costly off-chip DRAM accesses and optimise on-chip data reuse. An on-chip Global Buffer made with BRAM contains tiles of input feature maps and weights, which is used as an intermediate storage layer

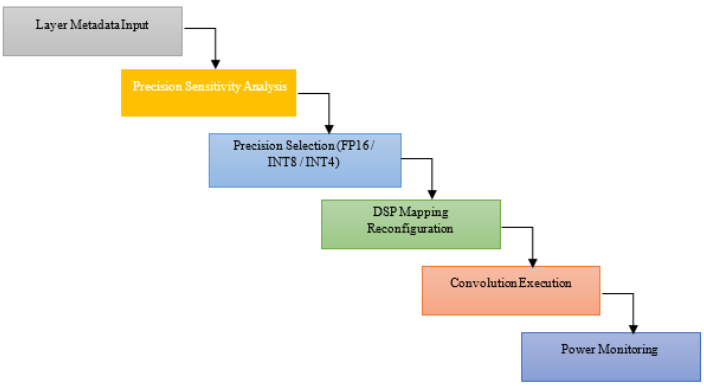

Fig. 2: Runtime Precision Control Unit (PCU) Flowchart for Dynamic Mixed-Precision Selection and DSP Reconfiguration

between data in external memory and the compute units. Local Processing Element (PE) Buffers are used to store weights and partial sums in temporary locations that allow retrieval in a short amount of time when doing convolution operations. The sliding window operations are supported by Line Buffers to provide efficient spatial data reuse through the storage of the overlapping rows of input feature map. Also, to implement a Double Buffering mechanism, data transfer and calculation phases overlap and one memory bank may be loaded with new data as the other is actively processing the current data. This hierarchical construction has a great way of lowering both the memory latency and energy usage by ensuring that most data migration is done in on-chip resources.

Power Elastic Scheduling System:

A runtime Power-Aware Scheduler evolves hardware activity dynamically as the workload intensity varies to improve further on the efficiency in terms of the consumed energy. The scheduler constantly cheques the computational load, the level of buffer occupancy and PE utilisation to compute the best operating conditions. According to this analysis, it varies the clock rate in order to trade off performance and energy usage. It also activates or deactivates compute clusters according to the real workload and eliminates idle processing units switching due to idle workload. This dynamic control means that hardware resources are utilised at the necessary capacity and this enhances energy proportionality in real-time inferences.

The strategy of Dynamic Power Management:

The energy optimization plan assimilates conditional power gating in order to remove idle losses of power. In situations where the use of computational resources reduces to a certain threshold, non-critical compute clusters are immediately shut down to avoid unnecessary switching transition. It is defined that the gating condition is.

$$\text{Utilization} < \text{Threshold}$$

where Utilisation is the percentage of used compute resource, and Threshold is an efficiency limit that is set by default. Once such condition has been met, modules are transitioned in low-power mode until the workload requirement goes up Figure 3. This hierarchical memory optimization and dynamic power control is synchronized to support low DRAM reliance, low power idle and high throughput to power ratio under the varying CNN workloads.

Results and Discussion

Experimental Setup

The suggested accelerator was deployed on a medium-level UltraScale+ platform powered by FPGA, in order to assess the performance of real-time CNN inference in edge deployment circumstances. The solid ResNet-18 and MobileNetV2, which are the most widely used benchmarks models, were chosen, as they exemplify the standard and lightweight CNNs. To provide realistic workload characteristics of classification, evaluation was performed using a subset of ImageNet dataset to ensure enough workload characteristics. The performance measurements were throughput in frames per second (FPS), overall power used in watts, energy efficiency in FPS/W and classification accuracy loss as compared to full-precision baselines. This experimental setup allows to conduct a thorough evaluation of computational capabilities, power scaling and precision scaling in conditions of real-time inference.

Performance Results

The experiment shows that the proposed architecture showed definite performance and energy efficiency gains.

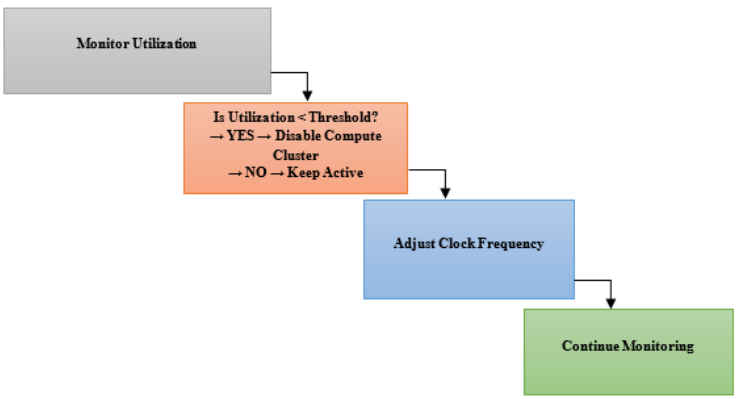

Fig. 3: Power-Aware Scheduling and Dynamic Compute Cluster Gating Based on Utilization Threshold

At 8.2 W, the accelerator reached 220 FPS in INT8 mode, which is 26.8 FPS/W with 1.2% accuracy loss at 8.2 W. In INT4 mode, throughput was higher at 260 FPS and power consumption was lower at 6.1 W and the resulting energy efficiency was much higher at 42.6 FPS/W, with a slightly large 2.8% error. MobileNetV2 run in INT8 mode used 7.4 W to get 310 FPS, which is equivalent to 41.8 FPS/W and a low loss of 0.9% accuracy. These findings prove that mixed-precision computing and adaptive dataflow optimization are efficient in enhancing throughput-per-watt without reducing classification performance, though, to a certain degree.

Comparative Analysis

Relative to a GPU-based embedded inference engine running in the same workload environment, the intended FPGA accelerator showed a difference in energy usage of about three times and an improvement in throughput-per-watt by 1.8 times Figure 4. This is largely due to slower movement of memory, effective then use of DSP and accuracy-sensitive computation. Moreover, the proposed design is 32% power savings, and 28% throughput more than a conventional fixed-precision FPGA accelerator. The following gains point to the benefits of using adaptive dataflow switching and runtime precision scaling with regards to fixed architectural designs.

Discussion

The findings suggest that memory hierarchy optimization can help save about 45 percent of energy consumption by inhibiting off-chip DRAM access by a substantial number. Dynamic mixed-precision execution can also further reduce switching activity, especially in later layers of the convolution where a less demanding numerical precision is needed. Along with that, the power-conscious scheduler avoids over-utilisation of computing resources by adjusting clock frequency dynamically and only enabling compute cluster when it is needed. Even though a small accuracy loss and minimal control overhead (less 2-percent) are introduced with INT4 mode and precision switching, respectively, the performance-energy trade-off is very attractive Table 1. A flexible combination of adaptive dataflow, Hierarchical memory management, dynamic precision scale, and workload - sensitive scheduling is significantly more effective than any of the optimization strategies applied alone, and this experiment illustrates the power of a holistic hardware-architecture co-design.

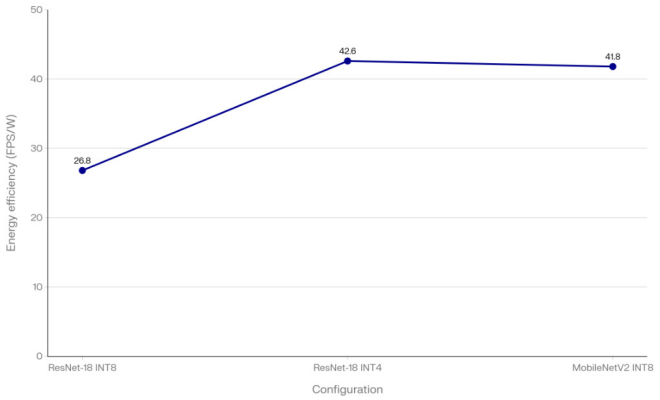

Figure 4: Energy Efficiency (FPS/W) Comparison Across Model Configurations and Precision Modes

Conclusion

This study introduces an innovative energy-saving FPGA-based accelerator interface with the specific purpose of the real-time CNN inference in edge computing systems. The proposed design is an extension to adaptive tiled dataflow to provide enhanced data reuse, dynamic mixed-precision computation to minimise switching activity and power usage, hierarchical on-chip memory optimization to minimise expensive DRAM accesses and power-aware scheduling to eliminate idle resource usage. The overall performance of the accelerator is achieved by harmonising these architectural methods in a shared paradigm where throughput-per-watts are significantly enhanced and classification accuracy with respect to representative CNN models are all in the competitive domain. The experimental analysis has shown that adaptive dataflow switching, precision scaling in run-time, and intelligent workload management can be used to achieve practical benefits compared to the traditional fixed-precision and fixed workload FPGA accelerators. The results validate that joint optimization of computation, memory hierarchy, and precision control are needed to provide scalable, high-performance, and energy-efficient solutions to deep learning inference to be implemented in next-generation edge AI systems.

Table 1: Performance and Energy Efficiency Evaluation of the Proposed FPGA Accelerator

Model / Configuration	Precision	Throughput (FPS)	Power (W)	Energy Efficiency (FPS/W)	Accuracy Drop (%)
ResNet-18	INT8	220	8.2	26.8	1.2
ResNet-18	INT4	260	6.1	42.6	2.8
MobileNetV2	INT8	310	7.4	41.8	0.9

References

1. Chen, Y., Zhu, L., Yu, L., & Yao, Y. (2020). Individual identification of communication radiation sources based on Inception and LSTM network. Journal of Physics: Conference Series, 1682(1), 012052. https://doi.org/10.1088/1742-6596/1682/1/012052
2. Ding, W., Huang, Z., Huang, Z., Tian, L., Wang, H., & Feng, S. (2019). Designing efficient accelerator of depthwise separable convolutional neural network on FPGA. Journal of Systems Architecture, 97, 278–286. https://doi.org/10.1016/j.sysarc.2019.04.005
3. Huang, J., Li, X., Wu, B., Wu, X., & Li, P. (2022). Few-shot radar emitter signal recognition based on attention-balanced prototypical network. Remote Sensing, 14(23), 6101. https://doi.org/10.3390/rs14236101
4. Li, H., Fan, X., Jiao, L., Cao, W., Zhou, X., & Wang, L. (2016). A high performance FPGA-based accelerator for large-scale convolutional neural networks. In 2016 26th International Conference on Field Programmable Logic and Applications (FPL) (pp. 1–8). IEEE. https://doi.org/10.1109/FPL.2016.7577308
5. Liu, Q., Han, L., Tan, R., Fan, H., Li, W., Zhu, H., & Liu, S. (2021). Hybrid attention based residual network for pansharpening. Remote Sensing, 13(10), 1962. https://doi.org/10.3390/rs13101962
6. Ma, Y., Cao, Y., Vrudhula, S., & Seo, J. S. (2017). An automatic RTL compiler for high-throughput FPGA implementation of diverse deep convolutional neural networks. In 2017 27th International Conference on Field Programmable Logic and Applications (FPL) (pp. 1–8). IEEE. https://doi.org/10.23919/FPL.2017.8056781
7. Mittal, S. (2020). A survey on modeling and improving reliability of DNN algorithms and accelerators. Journal of Systems Architecture, 104, 101689. https://doi.org/10.1016/j.sysarc.2019.101689
8. Pan, Y., Yang, S., Peng, H., Li, T., & Wang, W. (2019). Specific emitter identification based on deep residual networks. IEEE Access, 7, 54425–54434. https://doi.org/10.1109/ACCESS.2019.2913236
9. Sun, W., Wang, L., & Sun, S. (2021). Radar emitter individual identification based on convolutional neural network learning. Mathematical Problems in Engineering, 2021, 5341940. https:/ /doi.org/10.1155/2021/5341940
10. Wu, X., Ma, Y., Wang, M., & Wang, Z. (2021). A flexible and efficient FPGA accelerator for various large-scale and lightweight CNNs. IEEE Transactions on Circuits and Systems I: Regular Papers, 69(3), 1185–1198. https://doi.org/10.1109/TCSI.2021.3132702
11. Zhang, S., Pan, J., Han, Z., & Guo, L. (2021). Recognition of noisy radar emitter signals using a one-dimensional deep residual shrinkage network. Sensors, 21(23), 7973. https://doi.org/10.3390/s21237973