SCIENCE CHINA
Information Sciences

• LETTER •

March 2026, Vol. 69, Iss. 3, 139403:1–139403:2
https://doi.org/10.1007/s11432-025-4686-2

FPGA-based hardware accelerator designed for convolutional residual spiking neural networks

Yahui ZHANG[1,2], Shuiying XIANG[1,2*], Chenyang DU[1], Tao ZOU[1], Xingxing GUO[1,2], Ling ZHENG[3], Licun YU[4], Genquan HAN[2] & Yue HAO[2]

[1]*State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China*
[2]*State Key Discipline Laboratory of Wide Band gap Semiconductor Technology, School of Microelectronics, Xidian University, Xi'an 710071, China*
[3]*School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China*
[4]*CCCC First Highway Way Consultants Co. Ltd., Xi'an 710075, China*

Received 14 March 2025/Revised 28 May 2025/Accepted 18 November 2025/Published online 22 January 2026

Citation Zhang Y H, Xiang S Y, Du C Y, et al. FPGA-based hardware accelerator designed for convolutional residual spiking neural networks. Sci China Inf Sci, 2026, 69(3): 139403, https://doi.org/10.1007/s11432-025-4686-2

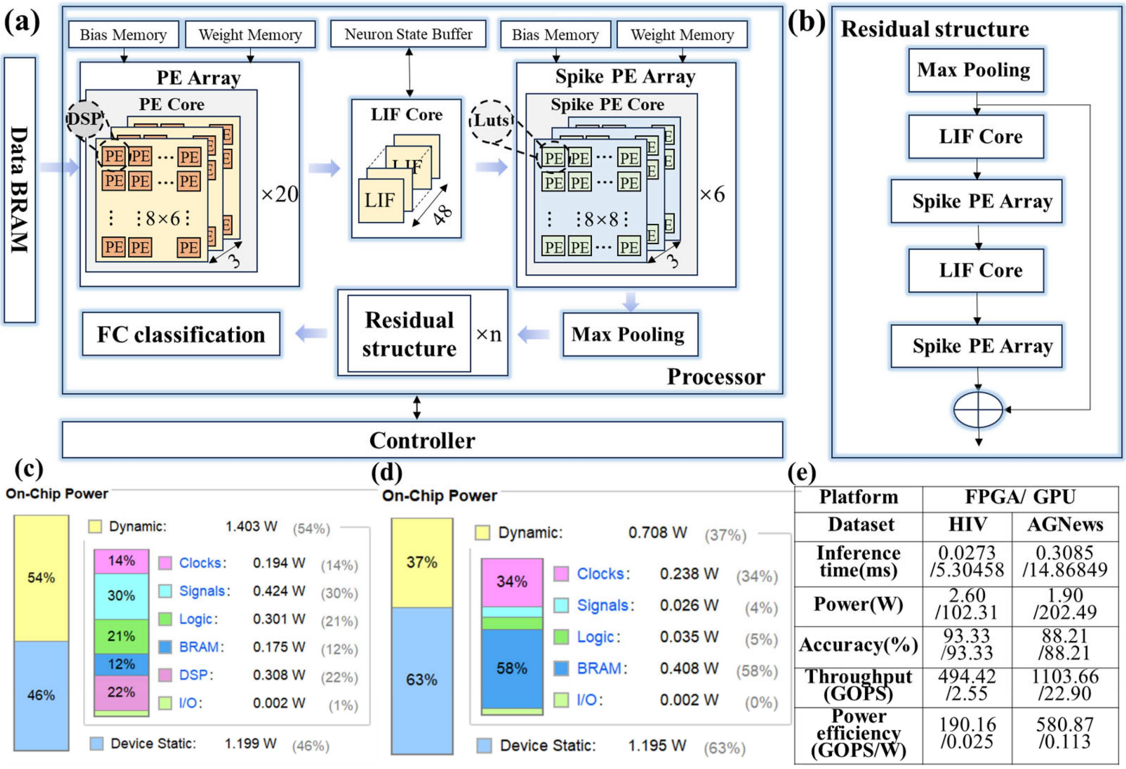
As the third-generation neural network models, spiking neural networks (SNNs) rely on discrete spatiotemporal spike sequences for information transmission and computation, featuring high brain biomimicry, energy efficiency, and low power consumption. These inherent advantages have enabled SNNs to be widely applied in diverse fields, including image and speech recognition, robotics control, and time series analysis. Currently, most SNN hardware platforms are CPUs and GPUs, both of which adopt the von Neumann architecture. This architecture separates processing units from memory, resulting in latency due to constant data exchange and, thus, reducing computational efficiency and throughput. However, field-programmable gate arrays (FPGAs) support direct interconnections between processing elements, delivering lower latency and higher energy efficiency after bitstream programming, while offering greater programmability and flexibility than ASICs. The deployment of FPGAs for SNN acceleration has already been investigated for several years. In 2025, a proposed convolutional neural network (CNN)-SNN hybrid accelerator reduced power consumption by 32% while maintaining 97.5% accuracy, improving FPS per watt by 47%–67% over conventional CNN architectures [1]. However, convolutional residual SNNs (CRSNNs) have not yet offered significant advantages on FPGA platforms. Herein, we deploy a CRSNN on an FPGA platform for gene analysis and text classification tasks. Results demonstrate that on both the HIV and AGNews datasets, the proposed implementation achieves reduced inference time and power consumption while maintaining the same accuracy as a GPU-based counterpart.

Design for FPGA-based accelerator. The architecture of the CRSNN is similar to that in [2]. The ZCU216 development board from Xilinx, equipped with the latest Zynq Ultrascale+ RFSOC 49DR main chip, is selected for deployment. The overall architecture of the ZCU216-based hardware accelerator is shown in Figure 1(a). The core components of the architecture are constituted by a processing element (PE) array, spike PE array, a leaky integrate-and-fire (LIF) core, and residual structure among others. The convolutional encoding of nonspiking signals in the spike encoder part is primarily undertaken by the PE array. These PEs are implemented using digital signal processors (DSPs). Twenty PE cores are employed in the spike encoder. Internally, each PE core is equipped with $3 \times 8 \times 6$ PEs. After multiplication in each clock cycle, adders are implemented. A total of $6 \times 1 \times 8 = 48$ adders are required per clock cycle. The spike neurons of the neuronal membrane potential across multiple time steps are implemented by the LIF core. The convolutional processing of spiking signals in the spike feature extractor part is handled by a spike PE array based on look-up tables (LUTs). On the HIV dataset, it is composed of six spike PE cores, with each core containing $3 \times 8 \times 8$ spike PEs. There are 48 parallel addition modules following multiplication modules, which are used to form a $6 \times 1 \times 8$ feature map. Other convolutional layers and the dataset follow a similar principle. All convolutions in the spike feature extractor part are achieved based on configurable logic block (CLB) LUTs and CLB registers. When the max-pooling layer is configured with a stride of 2, three comparisons are needed to generate the output for a single channel. The spiking residual structure used in this architecture is shown in Figure 1(b). The spiking residual network is composed of multiple spiking residual structure layers with spiking convolution. These residual connections are implemented using a combination of CLB LUTs and CLB registers. The fully connected classification layer performs further operations via multiplication and addition modules.

In the architecture, block random access memory (BRAM) is utilized to store network parameters and intermediate results, minimizing off-chip memory access.

Results. We programmed the bitstream file of the network onto the ZCU216 development board. An integrated logic analyzer was used to monitor the real-time computation results on the ZCU216. A frequency of 100 MHz was adopted in the ZCU216. The overall resource utilization of the CRSNN for a single sample in the HIV (AGNews) dataset included 24.21% (51.65%) of CLB LUTs, 6.17%

* Corresponding author (email: syxiang@xidian.edu.cn)



Platform	FPGA/ GPU	
Dataset	HIV	AGNews
Inference time(ms)	0.0273 /5.30458	0.3085 /14.86849
Power(W)	2.60 /102.31	1.90 /202.49
Accuracy(%)	93.33 /93.33	88.21 /88.21
Throughput (GOPS)	494.42 /2.55	1103.66 /22.90
Power efficiency (GOPS/W)	190.16 /0.025	580.87 /0.113

Figure 1 (Color online) (a) Overall architecture of a field-programmable gate array (FPGA)-based hardware accelerator designed for a convolutional residual spiking neural network (CRSNN); (b) architecture of the residual structure; on-chip power on the (c) HIV dataset and (d) AGNews dataset; (e) performance of the hardware accelerator.

(6.52%) of CLB registers, 67.42% of DSPs, and 3.3% (7.87%) of BRAMs of ZCU216. A pipelined approach was employed for processing multiple samples, allowing for efficient and continuous data processing. On-chip power consumption achieved on the HIV and AGNews datasets is shown in Figures 1(c) and (d).

Other performance metrics and comparisons with the Nvidia RTX3090 GPU model are shown in Figure 1(e). On the HIV dataset, the total required time was 0.0273 ms, meaning the FPGA-based inference time was 1/194.31 of that on the GPU platform. The total on-chip power was 1/39.35 of the power consumption of 102.31 W of the GPU platform. On the AGNews dataset, the inference time was 0.3085 ms, approximately 1/48.20 of that on the GPU platform. The power consumption reduced from 202.49 W on the GPU platform to 1.9 W on the FPGA, representing approximately 1/106.57 reduction. The accuracy, actual throughputs, and power efficiency on the two datasets are also shown in Figure 1(e). The performance was considerably enhanced on the FPGA platform, with the same inference accuracy as GPU. The inference speed of the proposed work outperformed those of state-of-the-art FPGA-based neural networks from other studies [1, 3].

Conclusion. We successfully deployed a CRSNN on an FPGA platform for gene analysis and text classification tasks. The evaluation was conducted using the ZCU216 board. On the HIV dataset, the FPGA-based inference achieved 0.0273 ms inference time and 2.6 W power consumption, representing approximately 194× speed-up and 39× reduction in power consumption relative to the GPU platform. On the AGNews dataset, the FPGA delivered 0.31 ms inference time and 1.9 W power consumption, accounting for approximately 1/48 of the inference time of the GPU platform and 1/106 of its power consumption. This deployment highlighted the potential of the FPGA-based accelerator for edge devices with constraints on time and power.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62204196, 62205258, U24B20137) and Fundamental Research Funds for the Central Universities (Grant No. QTZX23041).

References

1 Yun H, Park D. A power-efficient reconfigurable hybrid CNN-SNN accelerator for high performance AI applications. In: Proceedings of IEEE COOL CHIPS, 2025. 1–6
2 Zhang Y, Xiang S, Jiang S, et al. Hybrid photonic deep convolutional residual spiking neural networks for text classification. Opt Express, 2023, 31: 28489–28502
3 Gao Y, Wang T, Yang Y, et al. Advancing neuromorphic architecture towards emerging spiking neural network on FPGA. IEEE TCAD, 2025, 44: 3465–3478