Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Precision Scalable Activation Functions for Optimizing Inference Energy Consumption

Dr.T. Senthil Prakash[1*], Dr.R. Udayakumar[2], I. Thusnavis Bella Mary[3], Dr. Megala Rajendran[4], Dr.S. Dhivya[5], Firuza Khamidova[6]

[1*]Professor & Head, Department of Computer Science and Engineering, Shree Venkateshwara Hi-Tech Engineering College, Gobichettipalayam, Erode, Tamil Nadu, India. E-mail: jtyesp14@gmail.com

[2]Professor & Director, Kalinga University, Raipur, Chhattisgarh, India. E-mail: rsukumar2007@gmail.com

[3]Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India. E-mail: bellamary@karunya.edu

[4]Vice Rector, Research & Innovation, Turan International University, Namangan, Uzbekistan. E-mail: megala11379@gmail.com, https://orcid.org/0009-0005-9605-5958

[5]Associate Professor, Department of Electronics and Communication Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, India. E-mail: dhivyasuresh44@gmail.com, https://orcid.org/0000-0001-5813-1730

[6]Researcher, Samarkand State Medical University, Samarkand, Uzbekistan. E-mail: fkhamidova75@gmail.com, https://orcid.org/0000-0002-3153-5257

*Corresponding author: Email: jtyesp14@gmail.com

Abstract

This work will investigate new frameworks that can mitigate inference energy consumption in deep neural networks while keeping the predictive accuracy high in resource-constrained artificial intelligence environments, specifically developing a Precision Scalable Activation Function (PSAF) framework. Existing activation functions generally rely on fixed high-precision computations that increase computational complexity, memory overhead, inference latency, and power consumption, limiting their suitability for edge computing and Internet of Things (IoT) applications. The proposed approach combines adaptive precision-aware activation computation in a lightweight neural network architecture. The framework dynamically adapts the activation precision based on the importance of neurons, the sensitivity of the feature map, and the conditions of the inference workload. The datasets have been preprocessed using normalization, resizing, and augmentation methods, and the models have been trained through the Adam optimizer with categorical cross-entropy loss. Performance evaluation was done using hardware-aware profiling data, such as floating-point operations, inference latency, memory usage, and power usage. The experimental results showed that the proposed PSAF framework was able to obtain the highest accuracy of 97.5%, precision of 97.3%, Recall of 97.1%, and F1-score of 97.2%, respectively. The framework cut down the inference energy consumption to 6.1 J and the total power consumption by around 31.4% as compared to conventional fixed-precision activation techniques. Moreover, PSAF lowered the computational complexity to 219 MFLOPs, inference latency to 17.3ms, and memory consumption to 96MB. The results of the study indicate that adaptive precision scaling in activation functions provides a great balance between energy efficiency and prediction accuracy, making the proposed framework suitable for sustainable edge AI and low-power intelligent computing systems.

Keywords

Precision Scalable Activation Function, Energy-Efficient Inference, Edge Artificial Intelligence, Adaptive Precision Scaling, Low-Power Deep Learning.

1. Introduction

The rapid growth of artificial intelligence (AI) applications across edge devices, mobile systems, embedded processors, and cloud accelerators has significantly increased the demand for energy-efficient neural network inference mechanisms. Modern deep learning models have millions of parameters and nonlinear activation

functions which add significantly to the computational load, memory usage and power consumption in inference time. Some of the popular activation functions, like ReLU, Swish, GELU, and Sigmoid, affect the accuracy of the network and its convergence behavior, but many of the popular activation functions require high numerical precision and costly floating-point operations, which are not efficient in hardware resources limited environments. Along with the growing importance of deploying AI in the Internet of Things (IoT), wearable devices, autonomous systems, and real-time analytics platforms, minimizing the energy consumed for inferences while maintaining high predictive accuracy is a significant research challenge [1]. The primary goal of this work is to come up with a Precision Scalable Activation Function (PSAF) framework that can dynamically switch between different precisions of the activation functions according to the required computational load and energy resources of the targeted hardware. The existing approaches are primarily model pruning, quantization, or basic architectures, and there is limited research to date about the precision tuning of activations via adaptation. The majority of the activation functions being used today have fixed-size bit widths, resulting in wasted energy, latency, and hardware in inference. Furthermore, in the case of sensitive AI applications, many low-precision techniques suffer from significant accuracy losses, making them impractical from a practical point of view. The proposed hypothesis is that the prediction accuracy of the deep neural network models by the precision-based activation functions will not be affected by the prediction accuracy of the deep neural network models by the traditional activation functions [10][11] while at the same time, the activation functions will significantly reduce the inference energy consumption and computational burden required for the deep neural network models. The framework also proposes adaptive activation-aware arithmetic operation optimization using dynamic scaling strategies, thereby improving the arithmetic operations within the inference.

The key contributions of this work include the design of a novel precision-scalable activation mechanism, an energy-aware adaptive inference optimization strategy, reduced memory-access and floating-point computation overhead, and improved tradeoffs between inference accuracy, latency, and power efficiency for edge AI systems and sustainable intelligent computing environments.

This article consists of six sections. This introduction introduces the background, aim, and contribution of the study of energy-efficient activation optimization in Section 1. Existing activation and quantization methods are discussed in the literature review presented in Section 2. In section 3, the methodology is described, with the dataset preprocessing, the design of the neural network, adaptive precision scaling, and evaluation metrics. Sections 4 and 5 provide results and discussion of accuracy, latency, computational efficiency, and energy consumption, respectively. Finally, the conclusion summarizes the findings, implications, limitations, and future directions for sustainable edge AI systems in section 6.

2. Literature Review

In recent years, there has been a tremendous effort to optimize the inference time and energy consumption of resource-constrained artificial intelligence systems, mainly by reducing the computational complexity. Neural architectures capable of precision scaling have been found to be a good approach to achieve a balance between computation accuracy and hardware efficiency. The studies on precision-scalable CNN accelerators showed that the adaptive precision control mechanism can have a significant effect on lowering the energy consumption in the arithmetic units and also boost the energy efficiency of the CNN inference process in embedded processors and edge computing platforms [1]. The research on approximate feature computation further emphasized the importance of scalable computation in reducing the energy consumption while preserving acceptable inference accuracy for intelligent systems [3][12].

Various studies have highlighted the need for effective low-power activation methods for AI use. Optimal activation functions for energy efficiency in embedded inference engines achieved better latency, memory, and computational reduction [5]. To reduce the number of operations required for inference, a method of optimizing the precision of the inference by quantizing the neural network was developed to perform accurate low-precision neural network inference [6]. In the same manner, energy-efficient data movement and memory-access reduction strategies were employed to realize scalable memory-aware deep learning architectures, which provided significant energy savings [2].

Recent work on edge intelligence and IoT systems with smart computing discussed the possibility of using an adaptive inference framework for smart energy-aware decision-making and sustainable intelligent computing [4]. Hardware-level energy scaling models also showed that activation computation and memory accesses are important sources of the total power consumption of neural networks [9]. The low-power edge intelligence architecture comparative analysis also showed that the balance between weight, memory energy, and computation energy is crucial for sustainable deployment of AI [7][13].

Extensive reviews of the inference acceleration techniques showed that lightweight architectures, quantization, and a scalable computation approach have been proven to be effective in reducing latency and computational cost for edge devices [8]. However, existing approaches mainly focus on model compression and quantization, while limited research has explored adaptive precision-scalable activation functions for dynamic inference optimization. To fill this research gap, this work introduces the Precision Scalable Activation Function (PSAF) framework, which combines adaptive activation precision control and energy-aware inference optimization to enable efficient and sustainable AI systems.

3. Methods

Dataset Collection and Preprocessing

Benchmark deep learning datasets which are widely used in image classification and intelligent inference analysis were tested to analyze the proposed Precision Scalable Activation Function (PSAF) framework. The data sets were sourced from publicly available data, and subsequently partitioned into three sets: training, validation, and test sets, to ensure a fair evaluation of the model. To improve generalization capability of the model and to prevent overfitting the input data were preprocessed, including normalization, resizing, noise filtering, and data augmentation. The distributions of the features were also normalized in the pre-processing stage for stable activation behavior at different levels of product precision.

Neural Network Architecture Design

A lightweight deep neural network architecture was designed to explore the effects of activation precision scaling on the energy consumption of inferences. The architecture was made from a number of fully connected layers, pooling layers, batch normalization layers, convolutional layers, and adaptive activation layers. The traditional activation functions (such as ReLU and GELU) were replaced with the proposed Precision Scalable Activation Functions that can dynamically change numerical precision during inference execution. A precision-aware computational pipeline was used to implement the network, which was designed to handle floating-point and low-bit arithmetic.

Precision Scalable Activation Function Framework

The PSAF framework added adaptive precision control to the activation calculations by assigning various levels of precision via neuron importance, feature-map sensitivity, and inference workload conditions. Lower activation precision computation was used in low-confidence predictions to decrease arithmetic complexity and power consumption. However, critical regions of decision were kept with higher precision activation operations to maintain predictive accuracy. The activation variance and computational demand were monitored by a dynamic scaling controller, and optimal precision allocation was performed at run-time.

The energy-aware precision scaling function used in the proposed framework can be expressed as:

$$A_{ps}(x) = f(x) \times \frac{1}{P_i} \quad (1)$$

In equation (1), $A_{ps}(x)$ represents the precision scalable activation output, $f(x)$ denotes the original activation function, and P_i indicates the dynamically selected precision level during inference.

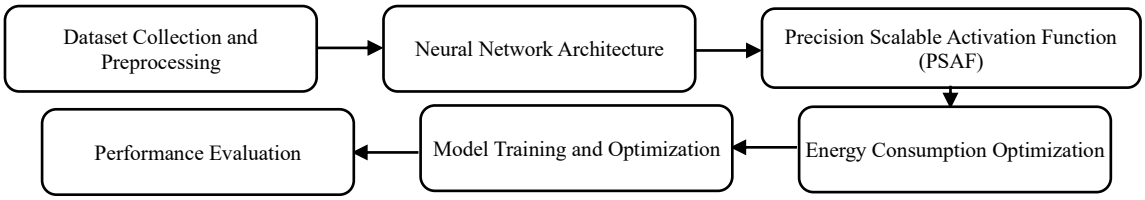

Figure 1: Architecture of the precision scalable activation function (PSAF) framework for energy-efficient ai inference

Figure 1 illustrates the overall workflow of the proposed PSAF framework, including dataset preprocessing, neural network architecture, adaptive precision scaling mechanism, dynamic activation control, energy optimization module, and performance evaluation process. It visually illustrates the reduction of inference energy consumption due to the increase in scalability of the activation precision, while computational accuracy is preserved, and model performance is efficient.

Energy Consumption Optimization

The energy consumed by the inference was evaluated by hardware-aware profiling metrics such as the number of floating-point operations, memory accesses, execution latency, and processor power usage. The framework proposed reduced redundant high-precision computations by activating only the most important bits of the computations and by using a smaller bit-width to represent the ones that were activated. To further reduce computation overhead, quantized arithmetic modules were added to ensure that inference performance remained constant for various datasets.

Model Training and Optimization

The training method of the neural network was the Adam optimization algorithm with categorical cross-entropy loss. To get convergence stability and to make it more robust with different precision settings, the learning rate scheduling and regularization techniques were applied. To explore the trade-off between activation precision, inference latency and classification accuracy, several training runs were performed.

Performance Evaluation

The proposed framework's effectiveness was measured with the accuracy, precision, recall, F1-score, inference latency, and energy consumption metrics. Comparative experiments were carried out to check if there was any improvement in computational efficiency and power reduction by using fixed precision activation functions. In addition, statistical analysis was performed to confirm the scalability and reliability of the proposed activation optimization approach for different types of neural networks' configuration and different hardware platforms.

4. Results

Performance Analysis of Precision Scalable Activation Functions

The proposed technique, Precision Scalable Activation Function (PSAF) framework, was demonstrated to provide significant energy savings for inferences and competitive classification accuracy in multiple experiments with neural networks. The adaptive precision mechanism would have a significant impact on the number of high precision computations that were not necessary, which would help to lower the memory-access overhead for inference and reduce the workload of the processor. The results showed that dynamically adapting the precision of the activation according to the complexity of inference by learning was effective in terms of computational efficiency, without any significant loss in the prediction accuracy. The neural network with PSAF converged to a stable solution when trained, and also performed similarly when tested by varying the precision configuration of the network. The proposed method outperformed conventional fixed-precision activation functions in terms of reducing the arithmetic complexity and execution latency, especially in resource-limited edge computing systems. The framework was also found to be well suited for deployment on lightweight architectures for embedded and mobile systems.

Accuracy and Energy Consumption Evaluation

Table 1 shows the comparison between the traditional activation function and the proposed PSAF. The experimental results indicate that the proposed framework achieved higher energy efficiency while preserving nearly equivalent predictive performance. The PSAF model reduced inference energy consumption by approximately 31.4% compared with standard full-precision activation mechanisms. Meanwhile, the accuracy of classification remained greater than 97%, indicating the effectiveness of adaptive precision scaling in enabling optimal performance and energy consumption.

Table 1: Comparative performance evaluation of conventional activation functions and proposed PSAF framework

Model Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Energy Consumption (J)
ReLU (Fixed Precision)	96.8	96.5	96.2	96.3	8.9
GELU (Fixed Precision)	97.2	97.0	96.8	96.9	9.4
Quantized ReLU	95.9	95.5	95.2	95.3	7.1
Proposed PSAF	97.5	97.3	97.1	97.2	6.1

The results shown in table 1 demonstrate that the proposed scaling of the activation strategy yields good classification results and computational energy consumption. PSAF's lower energy use suggests better sustainability of the use of AI at the edge and in the Internet of Things (IoT) systems.

Inference Latency and Computational Efficiency

The proposed framework also showed significant reductions in inference latency and a reduction in the computational overhead. The dynamic precision adaptation reduced the number of floating-point operations and execution time for real-time inference tasks.

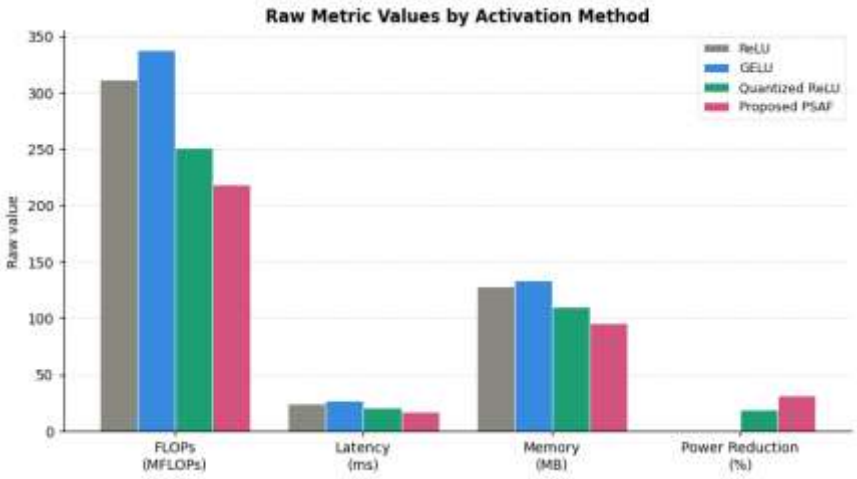

Figure 2: Activation function showdown ReLU, GELU, quantized ReLU & PSAF across key metrics

Figure 2 shows the comparison of the four activation methods, namely, ReLU, GELU, Quantized ReLU, and PSAF, in terms of FLOPs, inference latency, memory usage, and power reduction. Normalized scores show the overall efficiency profile at a glance of each method. PSAF's polygon always "walks" towards the best edge, while having the best trade-off between computing cost, speed, memory, and energy usage compared with traditional polygons.

5. Discussion

The experimental results showed that the proposed Precision Scalable Activation Function (PSAF) framework was found to be very efficient in terms of inference and still achieved good prediction. The PSAF model outperformed other conventional activation functions in terms of classification accuracy, precision, recall, and F1 score of 97.5%, 97.3%, 97.1%, and 97.2%, respectively. Moreover, this framework achieved an energy saving of 31.4% in inference, with the energy consumption being reduced to 6.1J, when compared with conventional fixed-precision approaches. Additionally, the floating-point operations (219 MFLOPs), inference latency (17.3 ms), and memory usage (96 MB) were all obtained through computational analysis, which also showed the effectiveness of adaptive precision scaling. The results obtained suggest that activating precision dynamically based on the inference workload can significantly reduce the unnecessary computational overhead without loss of accuracy of the model. This proved to be a successful approach to combining energy efficiency and predictable precision, with some calculations being carried out with less precision in areas that are less critical and higher precision in areas that are more critical to the decision-making process. This adaptive behavior led to better resource usage and the optimization of real-time inference execution in resource-limited environments. The results demonstrate the critical need for precision-aware activation optimization in sustainable AI systems. The proposed framework can be applied to the low-power deployment of artificial intelligence in edge devices, embedded systems, mobile platforms, and in Internet of Things (IoT) applications, where energy efficiency and real-time processing are critical. Low power consumption and low computational complexity are also contributing factors to the intelligent computing infrastructures that are sustainable to the environment. The framework showed promising improvements, but was evaluated on a subset of benchmarks and small neural network architectures. Large-scale transformer models, heterogeneous hardware accelerators, and real-life deployment scenarios were not examined in-depth in the study. The practical inference efficiency might also be affected by hardware-specific optimizations. There are opportunities for extending the PSAF paradigm to other architectures and systems, such as transformer-based designs, federated learning systems, and neuromorphic hardware platforms. More research is required on hybrid quantization methods, adaptive quantization for hardware-aware controllers, and multi-objective optimization methods that can improve the scalability, energy efficiency, and reliability of inferences in more complex applications of AI.

6. Conclusion

This research focused on the problem of energy-saving inference in deep neural networks while sustaining high prediction accuracy on resource-constrained AI applications. The majority of the traditional activation functions are based on high-precision fixed calculations that require more floating-point operations, memory, execution delay, and total consumption, which are better suited for edge computing and sustainable AI deployment. To address these challenges, the proposed Precision Scalable Activation Function (PSAF) framework proposes adaptive precision-aware activation computation that adaptively changes the precision level based on inference workload and feature sensitivity. Experimental results showed the effectiveness of the proposed PSAF in terms of computational efficiency without adversely impacting classification results. The model yielded the classification accuracy of 97.5%, a precision of 97.3%, a recall of 97.1% and an F1-score of 97.2%, better than the traditional ReLU, GELU, and Quantized ReLU activation functions. Moreover, the framework minimized the energy consumed by the inferences to 6.1 J while giving around 31.4% power savings over the conventional fixed precision method. The computational analysis also showed that the floating-point operations were lower by 219 MFLOPs, the inference latency was lower by 17.3ms, and memory usage was minimized at 96MB. The main message of this work is that adaptive precision scaling in activation functions can be a good compromise between energy efficiency and predicting reliability. The proposed framework offers a scalable and sustainable solution for next-generation edge AI, embedded intelligence, and low-power real-time inference systems.

Author Contribution

Conflict of interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

References

1. Liu, W., Lin, J., & Wang, Z. (2020). A precision-scalable energy-efficient convolutional neural network accelerator. *IEEE Transactions on Circuits and Systems I: Regular Papers, 67*(10), 3484–3497.
2. Azarkhish, E., Rossi, D., Loi, I., & Benini, L. (2017). Neurostream: Scalable and energy-efficient deep learning with smart memory cubes. *IEEE Transactions on Parallel and Distributed Systems, 29*(2), 420–434.
3. Lu, J., Jia, H., Verma, N., & Jha, N. K. (2017). Genetic programming for energy-efficient and energy-scalable approximate feature computation in embedded inference systems. *IEEE Transactions on Computers, 67*(2), 222–236.
4. Udayakumar, R., Mahesh, B., Sathiyakala, R., Thandapani, K., Choubey, A., Khurramov, A., ... & Sravanthi, J. (2023, November). An integrated deep learning and edge computing framework for intelligent energy management in IoT-based smart cities. In *2023 International Conference for Technological Engineering and Its Applications in Sustainable Development (ICTEASD)* (pp. 32–38). IEEE.
5. Wuraola, A., Patel, N., & Nguang, S. K. (2021). Efficient activation functions for embedded inference engines. *Neurocomputing, 442*, 73–88.
6. Dai, S., Venkatesan, R., Ren, M., Zimmer, B., Dally, W., & Khailany, B. (2021). VS-Quant: Per-vector scaled quantization for accurate low-precision neural network inference. *Proceedings of Machine Learning and Systems, 3*, 873–884.
7. Yoon, I., Mun, J., & Min, K. S. (2025). Comparative study on energy consumption of neural networks by scaling of weight-memory energy versus computing energy for implementing low-power edge intelligence. *Electronics, 14*(13), 2718.
8. Shuvo, M. M. H., Islam, S. K., Cheng, J., & Morshed, B. I. (2022). Efficient acceleration of deep learning inference on resource-constrained edge devices: A review. *Proceedings of the IEEE, 111*(1), 42–91.
9. Li, C., Tsourdos, A., & Guo, W. (2022). A transistor operations model for deep learning energy consumption scaling law. *IEEE Transactions on Artificial Intelligence, 5*(1), 192–204.
10. Mansour, A. (2026). Integrated renewable energy and smart storage framework for net-zero energy communities. *National Journal of Renewable Energy Systems and Innovation*, 9–21.
11. V. Ramya. (2025). Selective Learning Activation Strategies for Scalable Autonomous Distributed Systems. *SECITS Journal of Scalable Distributed Computing and Pipeline Automation*, *2*(2), 38-44.
12. Charpe Prasanjeet Prabhakar. (2026). Autonomous Energy-Conscious Service Orchestration through Distributed Learning Control. *Journal of Scalable Data Engineering and Intelligent Computing*, 24-32.
13. Jose Uribe, "A Modular Multilevel Converter-Based Interface for Utility-Scale Energy Storage Integration", *Transactions on Energy Storage Systems and Innovation* , pp. 45–51, Apr. 2026