Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A Reconfigurable Piecewise Linear Hardware Architecture for Efficient Floating-Point Division and Square Root Operations in High-Precision Softmax Accelerators

G. Kohila
M.E., Assistant Professor/ECE, Salem College of Engineering and Technology, Salem.

L. Aakash, D. Dhayaniyhi, P. Praveenkumar, A. Kaviyarasan
Salem College of Engineering and Technology, Salem.

Abstract - The rise of deep learning models, the Softmax function with self-attention has become integral in various real-time applications. Key operations in Softmax, such as floating-point division and square root, significantly affect computational accuracy but are costly in hardware, leading to high power consumption and large area requirements.These challenges hinder efficient deployment in latency-sensitive and resource-areconstrained environments. To address this, we propose a hardwareefficient piecewise linear approximation (PWL) unit that performs floating-point division and square root operations with low latency and reduced resource usage. By dividing the input range linear segments, precomputing slopes and intercepts, and using segment selection logic, the PWL unit approximates these nonlinear functions through simple multiplication and addition. Sharing the same unit for both operations further reduces area andpower consumption. Implementation results demonstrate that the proposed PWL-based design achieves high throughput and acceptable accuracy, making it well- suited for real-time deep learning tasks such as Softmax computation in self-attentionnetworks, while significantly reducing hardware complexity and energy consumption

INTRODUCTION

Deep learning has transformed the landscape of artificial intelligence, enabling machines to perform complex tasks such as image recognition, natural language understanding, and autonomous navigation with remarkable accuracy. At the heart of many state-of-the-art models is the self-attention mechanism, which allows the model to dynamically focus on different parts of the input data based on relevance. This mechanism relies heavily on the Softmax function, a mathematical operation that converts a vector of raw scores into a normalized probability distribution. By emphasizing important features and downplaying less relevant ones, Softmax enhances the model's ability to capture contextual relationships in data. Consequently, Softmax has become an indispensable part of much deep learning architecture, including the widely used Transformer models. As deep learning models find their way into real-world, timesensitive applications — such as autonomous driving, real-time video analytics, augmented reality, and mobile AI — the demand for hardware accelerators capable of delivering fast and efficient inference has grown exponentially. In these applications, low latency, high throughput, and energy efficiency are as critical as computational accuracy. However, the floating-point division and square root operations required by Softmax and other related functions are inherently costly when implemented on hardware. They typically demand iterative algorithms or complex logic circuits that consume significant silicon area and power. Furthermore, these operations often become bottlenecks in the computational pipeline, limiting overall system performance.

PROPOSED SYSTEM

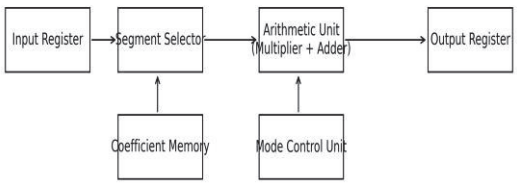
To address these challenges, this thesis proposes a novel piecewise linearapproximation (PWL) based hardware unit that performs floating-point division2 square root operations efficiently. The fundamental idea behind PWL is to approximate nonlinear functions by breaking their input domain into multiple small segments and representing each segment with a simple linear equation of the form: y=m×x+c where m is the slope and c is the intercept of the linear segment. By precomputing and storing the slope and intercept for each segment, the hardware can quickly approximate complex functions using just multiplication and addition operations. This drastically reduces computational complexity compared to iterative algorithms.In this design, the input range is divided into multiple intervals, and a segment selection logic determines which linear segment corresponds to the current input. The appropriate slope and intercept values are then retrieved from a lookup table, enabling fast computation of the approximate result. Importantly, the same PWL hardware unit is shared for both division and square

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

root operations by controlling input multiplexers and output logic, which further reduces silicon area and power consumption.This approach achieves a low-latency, low-power, and area-efficient hardware implementation suitable for real-time deep learning applications. While the approximation introduces some error, the granularity of the segments can be adjusted to meet the accuracy requirements of specific tasks, ensuring acceptable performance for Softmax computation within self-attention networks.The design and implementation of a reconfigurable PWL hardware unit capable of performing both floating-point division and square root operations. Development of an efficient segment selection mechanism and precipitation of slope and intercept values to enable fast and accurate function approximation. Detailedanalysis of the trade-offs between approximation accuracy, hardware complexity, and performance in the context of real-time Softmax computation.

proposed unit's effectiveness in reducing power consumption and silicon area while maintaining throughput suitable for latencysensitive deep learning workloads. Comparison with traditional non-restoring iterative methods and industry-standard IP cores, demonstrating significant improvements in hardware resource utilization and energy efficiency. The remainder of this thesis is organized as follows: Chapter 2: Literature Review –
Surveys existing algorithms and hardware implementations for floating-point
division,square root, and approximation techniques.

BLOCK DIAGRAM



CONCLUSION

The proposed system for hardware-efficient Softmax computation successfully addresses the limitations of existing designs by implementing a piecewise linear
approximation (PWL) approach for floatingpoint division and square root operations. Unlike conventional iterative methods, the PWL-based design significantly reduces latency, power consumption, and hardware complexity, making it highly suitable for real-time deep learning applications such as self-attention networks and transformer models. By dividing the input range into multiple linear segments and precomputing slope and intercept values, the system achieves high accuracy while maintaining low computational overhead. The use of a reconfigurable unit allows the same hardware to perform both division and square root operations, optimizing resource utilization and reducing silicon area. Pipelining and segment selection logic further enhance throughput, ensuring that the system can meet real-time performance requirements. The architecture's modular design and efficient arithmetic units demonstrate that approximate methods like PWL can provide a practical trade-off between accuracy and hardware efficiency. The system also supports flexibility for future extensions, such as higher-precision computations or integration with larger deep learning accelerators. Overall, the proposed design confirms that hardware-efficient approximation techniques can effectively replace traditional iterative methods for floating-point operations, achieving both high performance and energy efficiency. This work lays the foundation for future research in low-latency, resourceoptimized architectures for deep learning applications.

ACKNOWLEGMENT

This author would like to thank for support.

REFERENCES

[1] B. Liu et al., "Layer-wise mixed-modes CNN processing architecture with double-stationary dataflow and dimension-reshape strategy," IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 71, no. 10, pp. 4652–4664, Oct. 2024

[2] J. Kim, S. Kim, K. Choi, and I.-C. Park, "Hardware-efficient SoftMax architecture with bitwise exponentiation and reciprocal calculation," IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 71, no. 10, pp. 4574–4585, Oct. 2024

[3] Y. Fu, C. Zhou, T. Huang, E. Han, Y. He, and H. Jiao, "SoftAct: A highprecision softmax architecture for transformers supporting nonlinear functions," IEEE Trans. Circuits Syst. Video Technol., vol. 34, no. 9, pp. 8912–8923, Sep. 2024

[4] G. D. Meo, A. G. M. Strollo, D. D. Caro, L. Tegazzini, and E. Napoli, "Low-power high precision floating-point divider with bidimensional linear approximation," IEEE Trans. Circuits Syst. I, Reg. Papers, early access, Aug. 27, 2024

[5] G. Di Meo, A. Giuseppe Maria Strollo, and D. De Caro, "Novel lowpower floating-point divider with linear approximation and minimum mean relative error," IEEE Trans. Circuits
Syst. I, Reg. Papers, vol. 70, no. 12, pp. 5275–5288, Dec. 2023