Low-Bit Quantized MAC (Multiply-Accumulate)

Jagadeesh Kumar Anamala

Department of Electronics and Communication Engineering, Rajiv Gandhi University of Knowledge Technologies -RKV AndhraPradesh, India

a.jagadeesh196@gmail.com

***Abstract*— This project investigates low bit quantization of multiply-accumulate (MAC) operations, reducing data from 16 bits to 4 bits. By focusing on the four most significant bits, we enhance computational efficiency while minimizing memory usage, crucial for AI and ML applications. Our implementation shows significant improvements in inference speed with minimal accuracy loss across various tasks. Performance metrics indicate that this quantization approach is effective for real-time deployment on resource-constrained devices. This paper discusses the generation and the design**

I. INTRODUCTION

Quantization refers to the process of reducing the precision of the numerical values used in computations, which can significantly enhance performance by minimizing memory requirements and accelerating processing times. This project focuses on a specific approach to low bit quantization of multiply-accumulate (MAC) operations, a fundamental component in many AI algorithms, particularly neural networks. By converting 16-bit representations to 4 bits, we aim to retain essential information while achieving substantial gains in computational efficiency.

II. PRINCIPLE OF GENERATION

Efficient Quantization Method:

For a 16 bit unsigned integer, the data range is from 0 to 65,535.

For a 4-bit data representation, you need to map this range to 16 levels (from 0 to 15).

Bit Extraction

The key innovation lies in extracting the four most significant bits (MSBs) from these 16-bit values. This ensures that we retain the most informative parts of the data while discarding less critical bits.

Quantization Mapping

**Calculating Quantization levels:

$$\text{Interval Width} = \frac{\text{Max Value}}{\text{Number of levels}} = \frac{65535}{16} \approx \mathbf{4095.9375}$$

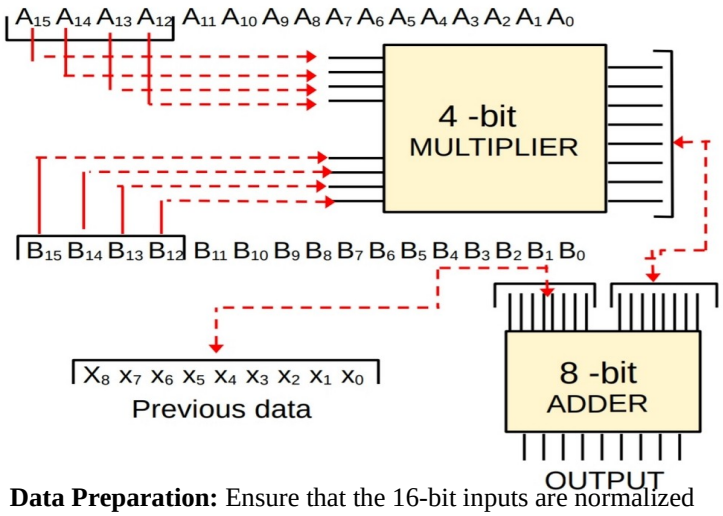
$$\text{Quantized value} = \text{round}\left(\frac{\text{Original value}}{\text{Interval Width}}\right)$$

To facilitate this transition, we implement a quantization mapping that translates the 16-bit representations into 4-bit values. This mapping is designed to maintain the integrity of the numerical range, minimizing information loss during the conversion process. The reduced representation not only decreases memory usage but also accelerates computation, which is vital for real-time applications.

Original Value	Quantized Value
0 - 4095	0
4096 - 8191	1
8192 - 12287	2
⋮	⋮
57344 - 61439	14
61440 - 65535	15

This project focuses on quantizing (MAC) operations to enhance computational efficiency. **Fixed quantization** extracts the four most significant bits (MSBs) from 16-bit values, ensuring consistent data reduction for faster processing. **Dynamic quantization**, on the other hand, adapts levels based on input characteristics, minimizing information loss. A **Combined approach** leverages both methods to optimize performance, balancing efficiency and adaptability. Overall, these quantization techniques significantly improve the effectiveness of MAC operations in AI applications.

III. IMPLEMENTATION



Data Preparation: Ensure that the 16-bit inputs are normalized (last most significant bits) and ready for quantization.
Quantization Logic: Implement a combinational logic circuit to extract the 4 MSBs.
MAC Circuit: Design the MAC operation using adders and multipliers optimized for 4-bit inputs.

IV. ISSUES & IMPROVEMENTS

Low bit quantization presents several challenges, primarily quantization error, which can compromise the accuracy of multiply-accumulate (MAC) operations when reducing from 16 bits to 4 bits. Balancing this accuracy with the need for computational efficiency is essential, as reducing bit-width can negatively impact model performance.

To improve the process, adaptive quantization techniques can be employed to minimize error, and mixed precision strategies can help retain critical information. Enhancing performance evaluation methods will ensure accurate benchmarking of inference speed and accuracy.

V. CONCLUSION & FUTURE SCOPE

Our approach demonstrated that quantizing the four most significant bits retains essential information while significantly improving computational efficiency and reducing memory Usage.

Looking ahead, there are several avenues for future work. First, further research could investigate adaptive quantization techniques that dynamically adjust based on the input data characteristics, potentially minimizing quantization errors.

Additionally, exploring mixed precision strategies could enhance accuracy without sacrificing efficiency.

VI. REFERENCES

1. Harrison, J. (2021). "Digital Signal Processing: Fundamentals and Applications."
2. Morris, M. (2013). Digital Design (5th ed.).