Efficient Edge-Based Inference Using Quantized Neural Networks

Abstract— We propose a 4-bit weight quantization method for CNNs, reducing memory usage by 68% with <3% accuracy loss.

Keywords

Edge AI; Quantization; Convolutional Networks; Embedded Systems.

Introduction

Edge devices benefit from reduced memory footprint and latency. We design QAT to preserve accuracy.

Results

Benchmarked on Raspberry Pi 5 with a 34% latency reduction and 68% memory savings.

Conclusion

4-bit quantization is viable for edge inference with minimal accuracy trade-offs.