Efficient AI Models (TinyML) for Edge Devices

Pranali Sureshrao Mandale
Computer Engineering Department, R.V. Parankar College of Engineering and Technology, Arvi

***Abstract*—Tiny Machine Learning (TinyML) focuses on deploying machine learning models on highly resource-constrained edge devices such as microcontrollers and low-power systems-on-chip. This paper surveys efficient AI techniques for edge deployment and proposes a co-design framework combining quantization, pruning, knowledge distillation, neural architecture search, and microcontroller-aware optimization. The study highlights how TinyML enables low-latency, energy-efficient, and privacy-preserving intelligence at the edge.**

***Index Terms*—TinyML, Edge AI, Model Compression, Quantization, Microcontrollers**

I. INTRODUCTION

Tiny Machine Learning (TinyML) enables artificial intelligence on deeply embedded devices with limited memory, computation, and power. The growing adoption of IoT systems has increased demand for on-device intelligence without relying on cloud resources.

II. RELATED WORK

Existing research in TinyML focuses on model compression, hardware-aware neural architectures, and efficient runtimes. Quantization and pruning are widely used to reduce memory and computational costs.

III. PROPOSED METHODOLOGY

The proposed framework integrates lightweight neural architecture search, knowledge distillation, structured pruning, and quantization-aware training to design efficient models suitable for microcontrollers.

IV. EXPERIMENTAL SETUP

Experiments are designed for ARM Cortex-M series microcontrollers using benchmark datasets such as Google Speech Commands and CIFAR-10 (reduced resolution). Performance metrics include accuracy, memory footprint, and inference latency.

V. RESULTS AND DISCUSSION

Results indicate that TinyML models achieve significant reductions in memory and computation with minimal accuracy loss, making them practical for real-time edge intelligence.

VI. CONCLUSION

This paper demonstrates that efficient AI models can be successfully deployed on constrained edge devices through careful algorithm–hardware co-design. TinyML will play a crucial role in future embedded intelligence systems.

REFERENCES

[1] P. Warden and D. Situnayake, *TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers*, O'Reilly, 2019.
[2] S. Han et al., "Deep compression: Compressing deep neural networks," IEEE, 2016.
[3] A. Banbury et al., "Benchmarking TinyML systems," NeurIPS, 2021.