Patent Search

Invention Title	Lightweight Machine Learning Model Compression for Real-Time Embedded Systems
Publication Number	15/2025
Publication Date	11/04/2025
Publication Type	INA
Application Number	202541032710
Application Filing Date	02/04/2025
Priority Number	
Priority Country	
Priority Date	
Field Of Invention	COMPUTER SCIENCE
Classification (IPC)	H04L0067120000, G06N0020000000, G06N0003063000, G06N0003045000, G06N0003082000

Inventor

Name	Address	Country
Karimunnisa Shaik	Assistant Professor IT Department Marri Laxman Reddy Institute of Technology and Management Hyderabad, Telangana.	India
K. Laxminarayanamma,	Head of the department, Computer Science and Engineering, Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana.	India
Yannam Apparao	Associate Professor, IT Department , Marri Laxman Reddy Institute of Technology and Management, Hyderabad, Telangana.	India
Asma Sultana	Assistant Professor, CSE Department, Marri Laxman Reddy Institute of Technology and Management, Hyderabad, Telangana.	India
B.Hemalatha,	Assistant professor, ECE Department, Anurag University, Venkatapur, Ghatkesar, Medchal-Malkajgiri District , 500088	India
Geetha Goli	Assistant professor, CSE Department, Marri Laxman Reddy Institute of Technology and Management, Hyderabad, Telangana	India
kestha Ravali	CSE Department, Marri Laxman Reddy Institute of Technology and Management, Hyderabad, Telangana	India

Applicant

Name	Address	Country
Marri Laxman Reddy Institute of Technology and Management	Marri Laxman Reddy Institute of Technology and Management Hyderabad, Telangana.	India
Anurag University	Anurag University, Venkatapur, Ghatkesar, Medchal-Malkajgiri District , 500088	India
Institute of Aeronautical Engineering	Computer Science and Engineering, Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana.	India
Karimunnisa Shaik	Assistant Professor IT Department Marri Laxman Reddy Institute of Technology and Management Hyderabad, Telangana.	India
K. Laxminarayanamma,	Head of the department, Computer Science and Engineering, Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana.	India
Yannam Apparao	Associate Professor, IT Department , Marri Laxman Reddy Institute of Technology and Management, Hyderabad, Telangana.	India
Asma Sultana	Assistant Professor, CSE Department, Marri Laxman Reddy Institute of Technology and Management, Hyderabad, Telangana.	India
B.Hemalatha,	Assistant professor, ECE Department, Anurag University, Venkatapur, Ghatkesar, Medchal-Malkajgiri District , 500088	India
Geetha Goli	Assistant professor, CSE Department, Marri Laxman Reddy Institute of Technology and Management, Hyderabad, Telangana	India

Abstract:

ABSTRACT [501] The invention relates to a Lightweight Machine Learning Model Compression Framework designed for real-time embedded systems, enabling efficier deployment of AI models on resource-constrained hardware. This innovation ensures optimized performance while reducing computational overhead, energy consur memory footprint. [510] By integrating quantization, pruning, and knowledge distillation techniques, the system effectively compresses machine learning models witl compromising accuracy, making them suitable for real-time execution on embedded devices, IoT systems, and edge computing platforms. [515] The framework empl adaptive compression strategy, dynamically adjusting the model size and computational complexity based on the available hardware resources, ensuring optimal per across various embedded environments. [520] Designed for low-power and real-time applications, the system minimizes latency and maximizes inference speed, ma for applications such as autonomous systems, real-time analytics, and smart sensor networks. [525] The invention integrates hardware-aware optimization technique seamless adaptation to different embedded architectures, including ARM processors, microcontrollers, and specialized AI accelerators, ensuring broad compatibility a embedded platforms. [530] The framework incorporates a hybrid compression approach that intelligently balances trade-offs between accuracy and efficiency, ensur performance in mission-critical applications such as automotive systems, industrial automation, and medical diagnostics. [535] The compressed models maintain hig accuracy and reliability, leveraging a fine-tuned training pipeline that preserves essential features while reducing unnecessary computational complexity. [540] The sy enables automated deployment and real-time model adaptation, allowing embedded AI applications to continuously optimize their performance based on real-worlc ensuring long-term efficiency and adaptability. [545] This invention revolutionizes AI deployment in embedded systems, providing a scalable, energy-efficient, and hig performance solution for real-time applications while addressing the growing demand for compact, power-efficient AI models in edge computing and IoT environmer

Complete Specification

Description:FIELD OF THE INVENTION

[501] The present invention relates to Lightweight Machine Learning Model Compression techniques specifically designed for real-time embedded systems. It addre growing need for deploying AI and machine learning (ML) models on resource-constrained devices such as IoT sensors, microcontrollers, edge computing platforms mobile processors.

[505] The invention is particularly useful in low-power, high-performance applications, where computational efficiency and reduced memory footprint are essential time inference. The system optimizes deep learning models through quantization, pruning, knowledge distillation, and neural architecture search to ensure lightwei deployment without compromising accuracy.

[510] The invention enables efficient model execution on embedded hardware architectures, including ARM Cortex, RISC-V, FPGA-based accelerators, and custom AI inference chips, ensuring compatibility across various industrial, healthcare, and automotive applications.

[515] By leveraging hardware-aware compression techniques, the framework ensures that machine learning models remain computationally efficient, power-optim suitable for real-time applications, such as autonomous vehicles, medical diagnostics, predictive maintenance, and smart home automation

View Application Status