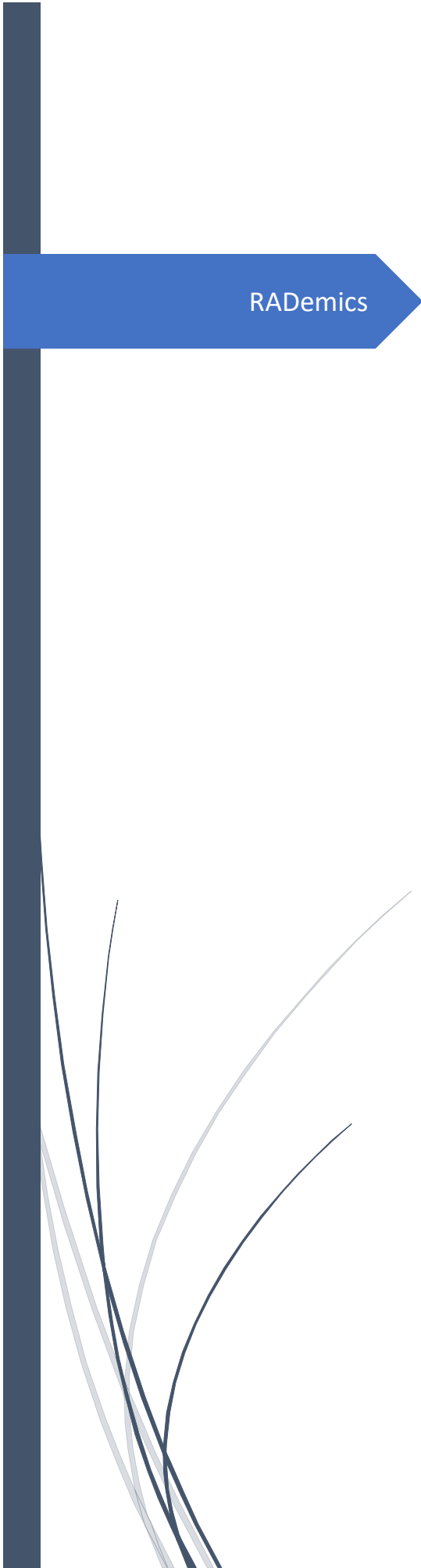


A thick dark blue vertical bar runs down the left side of the page. A blue arrow-shaped banner points to the right from this bar, containing the text 'RADemics'. Below the banner, several thin, curved lines in dark blue and light grey sweep upwards from the bottom left towards the center of the page.

RADemics

Integrating Edge Computing with Advanced Machine Learning Models in IoT

Manas Ranjan Mohapatra
BANKI AUTONOMOUS COLLEGE

14. Integrating Edge Computing with Advanced Machine Learning Models in IoT

Manas Ranjan Mohapatra, Lecturer, Department of Computer Science, Banki Autonomous College, Cuttack manasbanki@gmail.com

Abstract

The integration of edge computing with advanced machine learning models was transforming the landscape of IOT applications, driving significant advancements in real-time data processing and decision-making. This book chapter delves into the critical aspects of this integration, focusing on the optimization of machine learning models for edge environments, the development of edge-specific model architectures, and strategies for enhancing energy efficiency and reducing latency. Key areas of exploration include the deployment of lightweight models, leveraging hardware accelerators, and implementing flexible deployment strategies to address the constraints and requirements of edge devices. By examining the comparative performance of edge-specific versus traditional architectures, as well as benchmarking and evaluating model efficiency, this chapter provides a comprehensive framework for understanding and optimizing the interplay between edge computing and machine learning. The discussion was supported by practical case studies and real-world applications, offering valuable insights for researchers and practitioners seeking to enhance the capabilities and efficiency of edge-based IoT systems.

Keywords: Edge Computing, Machine Learning Optimization, Model Compression, Energy Efficiency, Latency Reduction, Edge-Specific Architectures.

Introduction

The rapid expansion of the IOT has led to a significant increase in the volume and complexity of data generated by interconnected devices [1]. As IoT applications become more prevalent, the demand for real-time data processing and analysis has intensified [2]. Traditional cloud-based computing models, while robust, often encounter challenges related to latency, bandwidth limitations, and scalability when handling vast quantities of data [3]. In response, edge computing has emerged as a paradigm shift that brings computational resources closer to the data source [4]. This proximity reduces latency and alleviates bandwidth issues, enabling more efficient and immediate data processing [5]. Integrating advanced machine learning models with edge computing offers the potential to address these challenges by enabling on-device analytics and decision-making [6].

One of the key challenges in integrating machine learning with edge computing was optimizing models to function effectively within the constraints of edge devices [7,8]. Unlike cloud-based environments with abundant computational resources, edge devices often have limited processing power, memory, and storage [9-11]. To address these constraints, machine learning models must be optimized for efficiency [12]. This includes techniques such as model compression, which reduces the size of models without significantly impacting accuracy, and quantization, which lowers the precision of computations to save on computational resources [13-15]. These

optimizations are critical for ensuring that models can operate effectively on resource-constrained edge devices while maintaining performance standards [16].

Energy efficiency was another crucial consideration when deploying machine learning models on edge devices [17]. Many edge devices are battery-powered or have limited energy resources, making it essential to minimize power consumption without compromising performance [18]. Techniques to enhance energy efficiency include optimizing algorithms to reduce computational complexity and leveraging hardware accelerators designed for low-power operation [19]. Additionally, efficient data handling and processing strategies can contribute to lower energy consumption [20]. By focusing on energy-efficient practices, the deployment of machine learning models on edge devices can become more sustainable, extending the operational lifespan of these devices and reducing the overall environmental impact.

Reducing latency was vital for applications that require real-time processing and instantaneous decision-making, such as autonomous vehicles or industrial automation systems [21-24]. Edge computing addresses latency issues by performing computations locally on edge devices, but further optimization was often necessary to achieve the desired performance. Techniques for latency reduction include algorithmic improvements, such as simplifying models and optimizing data pipelines, as well as utilizing specialized hardware for faster processing [25]. Benchmarking and evaluating these techniques are essential for ensuring that models meet the stringent latency requirements of edge-based applications, thereby enhancing their effectiveness and reliability.