International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Decentralized AI: The role of edge intelligence in next-gen computing

Dhruvitkumar V. Talati *

Independent Researcher, USA.

International Journal of Science and Research Archive, 2021, 02(01), 216-232

Publication history: Received on 10 February 2021; revised on 17 April 2021; accepted on 21 April 2021

Article DOI: https://doi.org/10.30574/ijsra.2021.2.1.0050

Abstract

With the rapid development of communication technology, the explosive growth of mobile and IoT devices, and growing requirements for real-time data processing, a new paradigm of computing, Edge Computing, has appeared. It moves computing power in the direction of data sources to mitigate latency, bandwidth usage, and dependence on cloud computing. In parallel, Artificial Intelligence (AI) has progressed notably with deep learning technology, highly optimized hardware, and distributed computing paradigms to yield smart applications of high computational loads. Nonetheless, the huge amounts of data generated on the network edge impose heavy challenges in managing data, network optimization, and implementing AI models. This has pushed the convergence of Edge Computing and AI, which has led to a new research area called Edge Intelligence.

Edge Intelligence is further divided into two broad categories

- AI for Edge (Intelligence-enabled Edge Computing) – This is concerned with augmenting Edge Computing architectures with AI-based methods, including resource management, task scheduling, computation offloading, and network optimization.
- Edge AI (Artificial Intelligence on Edge) – This involves executing AI models on edge devices directly, enabling local training and inference with minimal dependence on the cloud, thereby improving privacy, efficiency, and real-time processing.

This paper provides an overview of Edge Intelligence, including fundamental concepts, future technologies, and research directions. We identify critical challenges such as efficient deployment of AI models, decentralized AI learning through federated learning, and edge-centric accelerations of domain-specific hardware, and discuss how Edge Intelligence has the potential to transform domains like autonomous systems, smart cities, factory automation, and wireless networks. This paper documents the present trend and future path to act as the basis for researchers, engineers, and industry players seeking to improve the topic of AI-driven Edge Computing.

Keywords: Edge Intelligence; Edge Computing; Artificial Intelligence; Wireless Networks; Distributed AI; Computation Offloading; Federated Learning; Real-Time AI; Model Optimization; Ai Acceleration

1 Introduction

The Evolution of History in Communication Technologies and the Emergence of Edge Computing Communication technologies are evolving at a neck-breaking pace, with 5G networks leading the way to even swifter, more dependable, and highly efficient connectivity. The arrival of 5G not only improved network speed and bandwidth but also made it possible to introduce complex digital services, such as eMBB, mMTC, and uRLLC. These diverse service requirements are supported through end-to-end network slicing, a technique that virtualizes and segments network resources to provide tailored connectivity solutions for different applications.

* Corresponding author: Dhruvitkumar V. Talati.

At the center of this change lies the phenomenon of cloudification, with conventional hardware-based network functionality being shifted toward an end-to-end cloud administration model. Nevertheless, this shift extends beyond the centralization of cloud computing. Substantial amounts of the computing burden are increasingly being spread over regional cloud data centers and edge-cloud servers, bringing computation closer to end-users and mobile subscribers of services. This is due to the data explosion caused by Internet of Things (IoT) devices, which are becoming more ubiquitous across sectors.

By 2024, industry projections expect that close to 45% of internet data worldwide will be generated by IoT devices, which leads to gigantic amounts of data on the network edge. Conventional cloud architectures of computing are unable to work with such data optimally because of issues like network bottlenecks, latency, and bandwidth costs. It is not economical to transport vast amounts of data to central cloud servers for processing since it burdens the network infrastructure and adds inefficiency in real-time use.

Better is Edge Computing, a decentralized model of computing that processes data nearer to where it is created—on edge servers, gateways, and even end nodes—rather than solely depending on remote cloud platforms. Edge Computing is best applied to applications that demand ultra-low latency, high-bandwidth, and real-time responsiveness, including autonomous vehicles, smart cities, medical monitoring, and industrial automation. By bringing computational and storage power nearer to users, Edge Computing minimizes latency, maximizes effectiveness, and optimizes network utilization, thus becoming a key change in the existing computing infrastructure.

1.1 Artificial Intelligence: A Driving Force in Data-Driven Technologies

Along with the emergence of Edge Computing, Artificial Intelligence (AI) has been one of the most industry-transforming forces. Real-time processing and smart decision-making have fueled the need for AI-based solutions that can identify valuable insights from huge volumes of data.

Over the past decade, AI has seen an unprecedented explosion of breakthroughs, particularly with the advent of deep learning architectures. The success of AlexNet, Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) has driven innovations in computer vision, speech recognition, natural language processing (NLP), robotics, and predictive analytics. Deep learning algorithms excel at detecting intricate patterns in data, which has led to breakthroughs in autonomous systems, healthcare diagnostics, financial forecasting, and security analytics.

Besides the algorithmic improvement in performance, advancements in hardware have also contributed to expanding AI adoption. Specific computer hardware architectures, for example, Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), and other accelerators intended for AI purposes, have been optimized to provide high-performance computation for AI. Such hardware has contributed greatly towards improving throughput, efficiency, and scalability, thus speeding up AI model execution and power efficiency.

With increasing demand for real-time AI-based applications, Edge Computing and AI convergence, popularly referred to as Edge Intelligence, has become a significant research and development priority.

1.2 The Intersection of Edge Computing and AI: The Emergence of Edge Intelligence

As AI progresses and Edge Computing is picking up speed, the two are meeting to give rise to a new domain by the name of Edge Intelligence. Edge Intelligence is a combination of edge computation and AI-powered automation, decision-making, and prediction.

1.2.1 Edge Intelligence can be classified into two major categories

- AI for Edge (Intelligence-enabled Edge Computing): This aspect is concerned with the optimization of Edge Computing infrastructure through AI-driven methods. AI is applied to deal with critical issues surrounding resource management, task scheduling, network optimization, and edge security management. By incorporating AI in Edge Computing platforms, platforms can be self-optimizing, adaptive, and robust against changing environments.
- AI on Edge (Artificial Intelligence on Edge Devices): In this space, the whole AI model—training and inference—is deployed on the edge to minimize dependency on centralized cloud infrastructure. Execution of AI models on edge servers, IoT, and embedded systems enhances latency, security, and real-time decision-making. Methods like federated learning enable distributed training of AI models over edge nodes so that it can be feasible to provide privacy-preserving intelligence without raw data being moved to cloud data centers.

Edge Intelligence integration is transforming the majority of industries through real-time analytics, autonomous decision-making, and edge AI applications. Some of the major domains where Edge Intelligence is creating an impact are:

- Smart Cities: AI solutions through edges optimize traffic, surveillance, and city infrastructure to enhance efficiency and sustainability.
- Healthcare and Remote Monitoring: AI-based real-time edge analytics allow for continuous health monitoring, disease detection in the early stages, and tailored health recommendations.
- Autonomous Vehicles: Edge Intelligence supports low-latency decision-making for self-driving cars, enhancing navigation, collision avoidance, and vehicle-to-vehicle communications.
- Industrial IoT (IIoT) and Smart Factory: Edge analytics powered by AI supports predictive maintenance, production optimization, and quality inspection in smart factories.

The Scope of This Paper - This paper gives a complete overview of Edge Intelligence, its basic principles, research issues, and emerging trends. The paper is organized as follows

- Section II: Examines the building blocks and architectures of Edge Intelligence, including AI model deployment techniques and edge-native AI systems.
- Section III: Examines the contribution of computation offloading, federated learning, and distributed AI training to edge performance.
- Section IV: Examines challenges in Edge Intelligence, including resource constraint, security threats, model compression methods, and power-efficient AI inference.
- Section V: Explores current industry applications and use cases where Edge Intelligence is spearheading innovation.
- Section VI: Addresses directions for future research, such as AI-based edge security, real-time federated learning innovations, and future AI hardware acceleration.

By analyzing AI integration with Edge Computing, this paper will deliver the current situation and future direction of Edge Intelligence, providing a roadmap for developers, researchers, and industry stakeholders to extend the development of this revolutionary technology.

Table 1 Related surveys and their emphases

Perspectives	Related Surveys	Highlights
Intelligent Wireless Networking	[6] [7] [8] [9]	• Summarize the utilization of machine learning on the wireless edge • Including basic principles and general applications • Focus on resource management, networking, and mobility management • Optimization across different layers with machine learning technologies
Definitions and Divisions of Edge Intelligence	[10] [11] [12]	• Motivation, definition, division of Edge Intelligence • Including architectures, enabling technologies, learning frameworks, and software platforms • Focus on model training and inference on edge • Discuss the application scenarios and the practical implementations

1.3 Broadening the Horizons of Edge Intelligence: An Integrated Approach

The intersection of Artificial Intelligence (AI) and Edge Computing is not just technological convergence but also a paradigm that remakes real-time data processing, decision-making, and system optimization. With AI becoming functionally essential in large data analysis, insights creation, and automation drive, there is an increasing need to extend AI capabilities to edge environments. This has led to Edge Intelligence, which is an interdisciplinary area that enhances computational efficiency, responsiveness, and privacy-preserving AI at the edge.

However, Edge Intelligence is a far cry from the simple conjunction of AI and Edge Computing. It is an integrated, complex multi-faceted system involving distributed computing networks, network optimization based on AI, federated learning, and model deployment across resources. Despite its rapid progression, there still remains no collective definition of Edge Intelligence that everyone can agree to, with scholars and organizations continually expanding its sphere of influence and use.

1.3.1 Edge Intelligence: From Definitions to Contemporary Applications

Some researchers and research groups have sought to define and classify Edge Intelligence and provided differing views on its scope and model of architecture:

- Edge Intelligence and Synergistic Cloud-Edge Collaboration: Others believe that Edge Intelligence cannot be limited to isolated edge devices but should be executed in a collaborative setting between edge and cloud resources. In this method, cloud-edge collaboration provides scalable AI computation where computation workloads are dynamically assigned based on network conditions, latency demands, and device capabilities.
- Edge Intelligence and Autonomous AI: Some researchers see Edge Intelligence as a way to run full AI models on edge nodes that allow for on-device inference and training with no centralized cloud computing reliance. This definition reflects the vision of autonomous edge systems where AI algorithms are designed for low-resource environments and adaptive to live requirements.
- AI-Driven Fog Computing and Edge Optimization: Certain research advances typical Edge Computing by integrating Fog Computing paradigms in which AI informs network optimization, data caching, and adaptive model selection to improve spectral efficiency, bandwidth, and energy efficiency.

While these definitions show various aspects of Edge Intelligence, the area remains in its earliest stages of evolution, with growing interest from academics, businesses, and industry players who want to tap into real-time AI deployments in various industries.

1.4 Bridging the Gaps: A Structured Approach to Edge Intelligence

Although there is continuous research activity, most of the primitive terms of Edge Intelligence are vague and foremost challenges have to be untangled. The vast scope of Edge Intelligence entwines the majority of fields, such as network computing, AI model pruning, hardware accelerations, and data privacy, and it is hard to ascertain a unifying framework. This vagueness fueled the necessity for a more systematic approach of Edge Intelligence with bounded limits between its pieces and research directions.

To create a larger and more extensive vision, we suggest Edge Intelligence to be divided into two main domains

1.4.1 AI for Edge (Intelligence-Enabled Edge Computing - IEC)

This one is concerned with applying AI methods for Edge Computing infrastructure enhancement, addressing fundamental challenges of

- Resource Allocation and Task Scheduling: AI-enabled algorithms enhance the effectiveness of computational resources with efficient task offloading and network load balancing.
- Energy Efficiency Optimization: AI optimizes power management techniques, allowing low-energy AI inference and adaptive workload scaling on the edge device.
- Network Optimization and Adaptive Communication: AI-based solutions anticipate network congestion, adaptively scale bandwidth allocation, and optimize transmission of data, which enhances system performance.

Enabling Edge Computing infrastructure with AI, Intelligence-Enabled Edge Computing (IEC) empowers edge systems to self-optimize, evolve in response to changing circumstances, and provide high-performance computation in real time.

1.4.2 AI on Edge (Artificial Intelligence on Edge - AIE)

This category is engineered to execute and run AI models natively on edge devices, supporting:

- On-Device AI Inference and Model Training: AI models run locally on edge nodes, diminishing the use of centralized cloud infrastructures.
- Federated Learning for Decentralized AI Training: Rather than sending sensitive data to the cloud, federated learning allows edge devices to collectively train AI models in a privacy-preserving environment.
- Light and Efficient AI Models: Techniques like quantization, pruning, and distillation reduce the size of AI models to run faster on edge devices with limited resources.
- Real-Time AI: Applications like autonomous driving, industrial control, and health monitoring demand real-time AI insights, and therefore on-edge processing is crucial to real-time feedback.

Artificial Intelligence on Edge (AIE) ensures that AI capability is completely integrated into edge systems, enabling faster processing, lower data transfer cost, and increased privacy for mission-critical applications.

1.4.3 Motivation and Scope of This Study

While previous work has touched on some facets of Edge Intelligence briefly, the domain has not yet seen a unifying model delineating the distinction between AI-aided optimization of Edge Computing (AI for Edge) and AI deployment in accordance with Edge (AI on Edge). The aim of this research is to address such lacunae by presenting a thoroughly organized overview with observations into:

The interplay between Edge Computing and AI, specifying how their interaction gives rise to Edge Intelligence.

- A multi-tiered research agenda, defining most critical technological building blocks, architectural designs, and upcoming trends in Edge Intelligence.
- Future research challenges, including model effectiveness, security exposure, real-time adjustability, and scalable AI deployment.
- To that end, the paper is organized as follows:
- Section II: Discusses the intersection of Edge Computing and AI, defining their dependencies and complementarities.
- \\Section III: Reveals the Edge Intelligence research roadmap, dissecting its technological layers and deployment strategies.
- \\Section IV & Section V: Explore AI for Edge and AI on Edge, respectively, outlining their intrinsic principles, practical applications, and current issues.
- \\Section VI: Draws the study to a close by outlining future research directions and opportunities of Edge Intelligence development.

By presenting a systematic and well-defined taxonomy of Edge Intelligence, this article hopes to set the stage for future innovation, directing researchers and industry experts toward creating more intelligent, efficient, and scalable AI-powered edge ecosystems.

1.5 Conclusion: The Future of Edge Intelligence

Edge Intelligence is the next thing in AI-powered computing that redefines data processing, analysis, and usage in real-time scenarios. As companies increasingly seek low-latency and high-efficiency AI, the confluence of AI and Edge Computing will dictate future technologies.

With continuing innovation in federated learning, light AI models, edge-native hardware accelerators, and network optimization through AI, Edge Intelligence will be a major driver for smart systems for many applications in markets such as:

- Smart Cities and Infrastructure
- Healthcare and Wearable Technology
- Transportation Systems and Autonomous Vehicles
- Industrial Internet of Things and Smart Factory
- 5G and Next Generation Wireless Networks

As Edge Intelligence evolves, it will not only lead to greater computational power but also gigantic jumps in AI autonomy, security, and scalability, making it one of the most revolutionary and promising areas of modern computing.

The aim of this work is to act as a guide reference, making an overall vision of Edge Intelligence, preparing the ground for further researches and practical implementations in AI-powered Edge Computing.

The new breakthrough in Artificial Intelligence and Edge Computing has enabled Edge Intelligence, the revolutionizer that is making smart, high-performance, and real-time computing accessible to the network edges. With AI of low latency and high performance being increasingly utilized in industries, it will become imperative to integrate the AI with edges while designing smart systems of the future.

Since there is continuous innovation happening in AI model performance, edge-native hardware capabilities, and distributed intelligence architectures, Edge Intelligence holds the promise of redefining computing paradises with unprecedented edge automation, analysis, and decision-making opportunities.

This article provides a backdrop for comprehending the revolutionizing effect of Edge Intelligence by crossing the gap between cloud-based artificial intelligence and edge-decentralized computation, and paving the path for future technological innovations in AI-powered edge technology.

2 The Synergistic Combination of Edge Computing and AI

The intersection of Edge Computing and Artificial Intelligence (AI) is not coincidental but the natural and progression towards the future in the digital age. The two technologies are complementary, with AI augmenting the functionality of Edge Computing, and Edge Computing amplifying the reach and potential of AI. Combined, their ability to remake the landscape of real-time data processing, decision-making, and intelligent automation across industries cannot be overstated.

AI, in effect, secures Edge Computing by bringing in the capability of optimization, machine learning-based decision-making, and autonomous management of systems. Edge Computing, conversely, pushes the extent of AI through offering decentralization-based, low-latency environments with which AI can be deployed beyond the confines of centralized cloud computing spaces. Such syncretism delivers Edge Intelligence as a framework to make AI-based applications more efficiently, autonomously, and securely operate within decentralization contexts.

2.1 How AI Augments Edge Computing

Edge Computing is fundamentally a distributed paradigm for computing, where computing power is distributed to minimize latency, maximize scalability, and enhance system resilience. But distributed network management and real-time optimization of resource utilization are very critical challenges. AI addresses these problems by providing intelligent solutions to optimize system functioning, maximize efficiency, and facilitate autonomous decision-making at the edge.

A number of AI-powered methods enable Edge Computing in the following manners

2.1.1 Smart Resource Optimization and Allocation

- Edge Computing entails various levels of resource allocation, such as CPU cycle scheduling, bandwidth allocation, access control, and network frequency assignment.
- AI-powered optimization models—via methods such as stochastic gradient descent (SGD), statistical learning, and deep learning—can allocate and re-allocate these resources better than conventional rule-based systems.

2.1.2 Reinforcement Learning for Dynamic Edge Management:

- Multi-armed bandit theory, multi-agent learning, and deep Q-networks (DQN) reinforcement learning algorithms are being used more and more to address intricate decision-making problems in Edge Computing.
- Artificial intelligence-based adaptive learning mechanisms allow edge networks to adapt automatically, optimizing task scheduling, load balancing, and energy consumption without human intervention.

2.1.3 Autonomous Edge Operations Using AI Models

- AI systems can assist edge devices to forecast and diagnose failures, allocate computing resources dynamically, and optimize performance autonomously through real-time data streams.
- Deep learning models like CNNs and RNNs empower Edge Computing applications of image processing, speech processing, and real-time analysis.

- With the addition of AI-powered methods, Edge Computing becomes smart, adaptive, and skilled in performing elaborate, data-consuming tasks.

2.2 How Edge Computing Augments AI

While AI gives the computational intelligence to Edge Computing, Edge Computing gives infrastructure as well as grass-root applicability to AI technology. The mass use of Internet of Things (IoT) devices and Internet of Everything (IoE) has seen an unprecedented amount of data that traditional centralized cloud servers are unable to process effectively. Edge Computing offers AI the distributed environments in which AI models can process, analyze, and infer insights close to where the data is generated.

2.2.1 Real-World Use Case Scenarios for AI

- Increased usage of smart devices, self-driving cars, and IoT-connected infrastructure has developed explosive demands for real-time processing.
- Edge Computing makes it possible for AI-powered applications such as:
- Smart transportation systems (e.g., autonomous cars, intelligent traffic management).
- Smart cities (e.g., AI-enabled surveillance, weather forecasting).
- Health and public safety (e.g., real-time disease diagnosis, AI-enabled emergency response).
- Such uses require AI models to run autonomously at local environments, and hence Edge Computing is one of the key enablers of AI deployment outside the cloud.

2.2.2 A Decentralized Platform for Execution of AI Models

- Edge Computing offers a decentralized platform that enables AI models' run-time on edge servers, IoT gateways, and smartphones.
- These compute-intensity low, communication-need high AI use cases can be offloaded from edge to cloud and decrease reliance on central processing centers.

2.2.3 Edge AI Hardware Advances

- The advent of dedicated AI acceleration hardware enabled AI inference and training on edge devices with ease.
- Advances in Field Programmable Gate Arrays (FPGAs), Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), and Neural Processing Units (NPUs) chips have enabled edge devices to run complex AI models with negligible latency.
- Research organizations and vendors are now developing new chip architectures specifically for edge computing so that deep learning models can be deployed on low-power and resource-limited edge hardware.

By giving AI a decentralized computing environment, Edge Computing removes most of the usual bottlenecks of cloud-based AI, including network traffic, latency delays, and security loopholes.

2.3 The Future of AI-Edge Computing Convergence

As AI and Edge Computing continue to advance in the future, their integration will open the doors to smart, real-time, and decentralized computing environments. The confluence of AI and Edge Computing will be expected to:

2.3.1 Facilitate Ultra-Low Latency AI Applications

Applications based on AI like autonomous drones, industrial automation, and smart healthcare monitoring will enjoy real-time data processing at the edge without cloud communication that is time-consuming.

2.3.2 Improve AI Model Performance with Distributed Learning

Federated learning and distributed training of AI at the edge will enable devices to collaborate and assist in AI model training in a collective manner without exposing raw data, ensuring user privacy and data security.

2.3.3 Scale AI Deployment to Resource-Constrained Devices

Edge Computing will enable AI power-optimized models to run on battery-powered IoT devices, opening up the way for AI-powered wearables, sensors, and smart home assistants.

2.3.4 Drive Next-Generation Intelligent Networks

The synergy between network optimization via AI and Edge Computing infrastructure will facilitate the installation of 6G networks, real-time cybersecurity defense systems, and adaptive AI-driven edge frameworks.

The relationship between AI and Edge Computing is inherently symbiotic, with each technology enhancing the capabilities of the other. AI empowers Edge Computing by enabling intelligent resource allocation, adaptive learning, and automated decision-making, while Edge Computing expands AI s reach by providing real-world deployment environments, reducing latency, and improving efficiency.

As AI becomes increasingly decentralized and deployed in edge environments, Edge Computing will form the basis of next-generation intelligent applications, facilitating real-time AI inference, power-efficient machine learning, and distributed intelligence security.

The marriage of AI and Edge Computing—Edge Intelligence—will be a key enabler of next-gen computing, facilitating innovations in autonomous systems, industrial IoT, healthcare technology, smart city creation, and more.

2.3.5 This paper will proceed to explore

- Section III: Edge Intelligence's research roadmap, deciding on its technology layers and its future directions.
- Section IV & V: In-depth analysis of AI for Edge and AI on Edge, deciding on architectural breakthroughs, real-world applications, and shared issues.
- Section VI: The future direction of Edge Intelligence, focusing on future research domains and industry adoption plans.

By defining a clear and organized comprehension of the interconnection between AI and Edge Computing, this research seeks to provide the foundation for future innovations, leading researchers and technology innovators to create intelligent, scalable, and real-time AI-driven edge ecosystems.

3 Edge intelligence research roadmap

The Edge Intelligence research roadmap defines the architectural layers, major components, and optimization approaches that drive its development. As shown in Fig. 1, Edge Intelligence can be explored from two basic directions:

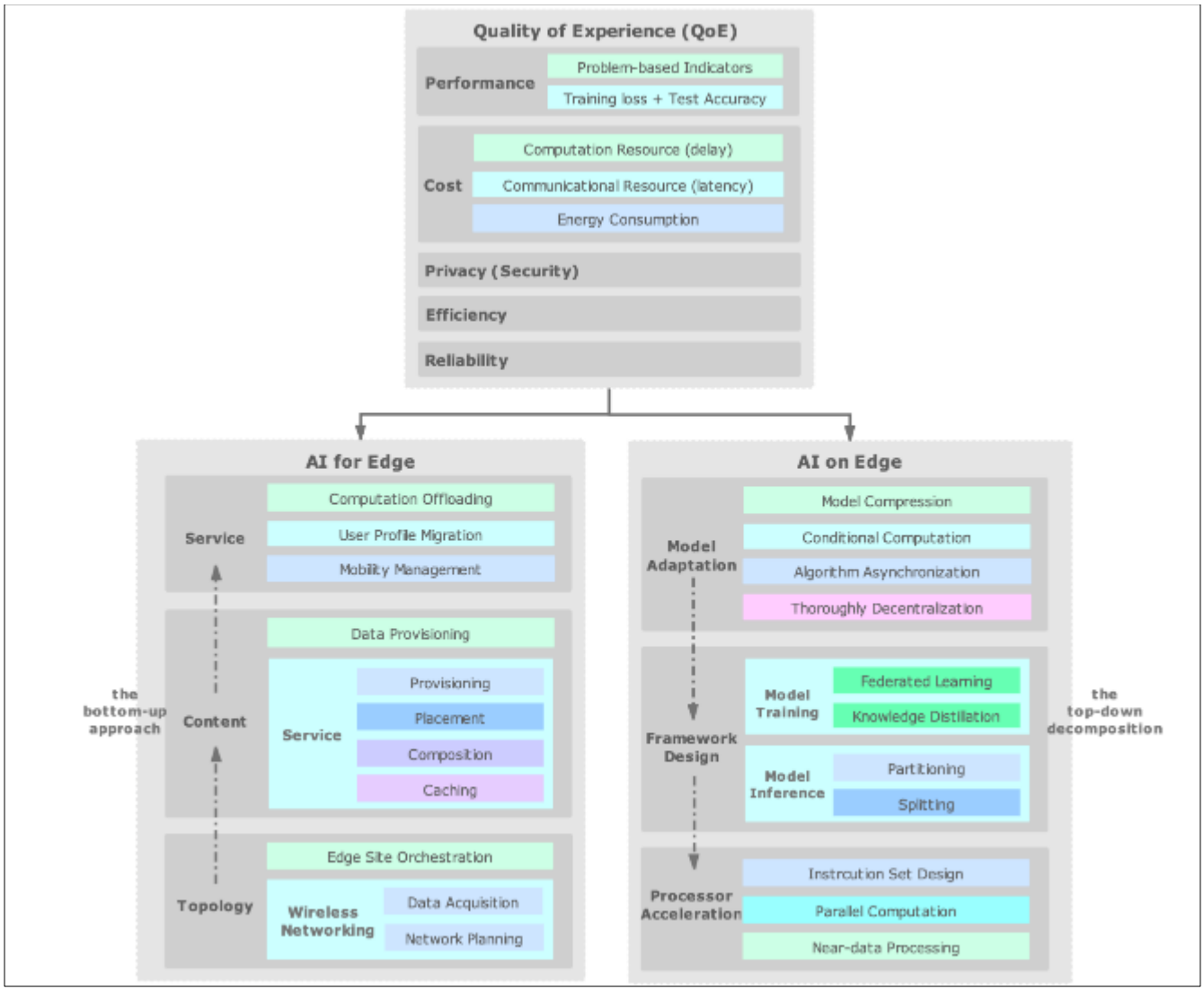

Figure 1 The research rod-map of edge Intelligence

- AI for Edge (Intelligence-Enabled Edge Computing): This path involves using AI methods to optimize Edge Computing platforms, solving issues such as resource allocation, network optimization, and energy efficiency.
- AI on Edge (Artificial Intelligence on Edge): It is centered around putting AI models on the edge itself, providing real-time processing, decentralized AI training, and privacy-preserving inference.

In order to study Edge Intelligence systematically, we break research activity into two complementary paradigms:

- Bottom-Up Approach: Researching Edge Computing layers, i.e., topology, content, and service—where AI can be utilized for optimization.
- Top-Down Decomposition: AI adaptation research in Edge environments, such as model optimization, framework design, and hardware acceleration.

Quality of Experience (QoE) is the general theme of this research roadmap since it is the main optimization goal for both AI for Edge and AI on Edge. QoE is characterized by five major factors: Performance, Cost, Privacy & Security, Efficiency, and Reliability.

3.1 Quality of Experience (QoE) in Edge Intelligence

Quality of Experience (QoE) is an application-dependent measure that assesses the efficiency and efficacy of AI-powered Edge Computing systems. It takes various parameters into account to make AI-fueled Edge Computing systems offer the optimal user experience.

3.1.1 Performance: Best Computational Efficiency

- Performance measures differ for AI for Edge and AI on Edge.
- For AI for Edge, performance is measured in terms of task accomplishment rates of execution, for example:
- Offloaded tasks from the edge devices successfully as a percentage to servers.
- Service providers' capability to achieve maximum revenue and minimum cost of resources (e.g., number of base stations required to maximize).
- In AI on Edge, performance is dependent on AI model efficiency, and these are:
- Training loss: Accuracy of training models on edge devices.
- Inference accuracy: Accuracy of models deployed on the edge to make correct predictions.
- Even with the shift from cloud-to-edge AI computation, these parameters continue to be at the heart of Edge Intelligence evaluation.

3.1.2 Cost: Minimizing Computational and Communication Overheads

- Three important elements make up the cost reduction of Edge Intelligence:
  - Computational Cost: CPU cycle requests, assigned processing time, and energy consumption of AI operations within the edge.
  - Communication Cost: Data communication bandwidth and network resources consumed for data transfer between edge nodes, gateways, and cloud servers.
  - Power Consumption: Because edge devices possess limited sources of power, efficient AI deployment needs to consume minimal battery with optimal performance.
- Edge Intelligence seeks to minimize latency and energy expenses as much as possible and enable possible real-time processing of AI on edge devices with no overhead.

3.1.3 Privacy & Security: Secure Data at the Edge

- Privacy issues in Edge Intelligence are caused by the sensitivity of user information that edge devices (e.g., smartphones, wearables, autonomous vehicles) capture.
- Federated Learning has come as a solution, enabling AI models to be trained locally on edge devices without communicating information directly to central servers.
- Security issues in Edge Intelligence are:
- Avoiding cyber-attacks on edge networks.
- Ensuring secure AI model deployment over heterogeneous edge infrastructures.
- Improving middleware resilience to avoid system vulnerabilities and misuse.
- Mitigation of these security threats is essential to make AI-driven Edge Computing ecosystems trustworthy.

3.1.4 Efficiency: Best AI Model Deployment on the Edge

Efficiency is at the core of both AI on Edge and AI for Edge, ensuring the most performance with no wastage.

The best AI-powered strategies for the most efficiency are

- Model Compression: Compressing the size and complexity of AI models so that they can operate on low-resource edge devices.
- Conditional Computation: Conditionally modulating AI computations according to the actual conditions at the edge.
- Algorithm Asynchronization: Parallel computation and distributed processing enablement of AI training workflows.
- These strategies make Edge AI applications more scalable and responsive and enable low-power devices to run AI models effectively.

3.1.5 Reliability: Guaranteeing Strong and Connected AI Execution

- Uninterrupted AI-driven operations at the edge must be supported by reliability in order to ensure smoothness.
- A few of the key reliability issues in AI on Edge are:
- Distributed fault-tolerant training: AI models trained on various edge nodes are plagued with incomplete updates due to network failures.
- Volatility of wireless networks: Edge AI models need to be hardened against network volatility, ruling out service interruption during model updating and data synchronization.
- Improved system reliability guarantees that AI models execute automatically in mission-critical edge applications like autonomous vehicles, industrial automation, and smart healthcare.

3.2 Edge Intelligence Research Dimensions

The Edge Intelligence research agenda is bifurcated into two broad directions:

3.2.1 AI for Edge: Intelligence-Enabling Edge Computing

It deals with the application of AI methods to enhance Edge Computing infrastructure, tackling issues in

- Computations Offloading: AI strategies decide what operations should be carried out locally and offloaded on distant servers.
- Mobility Management: AI-driven Edge systems dynamically handle mobile users' and networked vehicle service provision.
- Resource Optimization: Machine learning optimizes bandwidth use, computation capability, and storage between edge nodes.
- Task Content Cache and Data Streaming: AI facilitates sophisticated real-time data streaming and anticipatory caching to generate enhanced efficiency.

3.2.2 AI on Edge: Artificial Intelligence on Edge Devices

Here, emphasis is placed upon running AI models directly on the edge device for the following benefits

- Federated Learning: Learning AI on diverse edge devices in a way that protects user privacy.
- Knowledge Distillation: Knowledge from big cloud-trained models is distilled and transferred to tiny AI models on edge devices.
- Model Inference Optimization: AI models are optimized to support low-latency real-time decision at the edge.
- Processor Acceleration: Hardware acceleration in edge-focused AI chips (e.g., NPUs, TPUs, FPGAs, GPUs) accelerates AI performance and energy efficiency.

3.3 The Future of Edge Intelligence Research

Edge Intelligence is a constantly advancing area with research challenges and technological innovation that are constantly increasing. The AI-driven Edge Computing future will be centered on:

3.3.1 Ultra-Efficient AI Models for Edge Devices

- Promoting low-energy deep learning models that are deployable on edge devices.
- Edge-native AI systems that are optimized for real-time inference with zero or negligible energy cost.

3.3.2 Advanced Federated Learning and Decentralized AI:

- Evaluating privacy-preserving AI training in heterogeneous edge environments.
- Data synchronization protocol optimization to address communication bottlenecks in federated learning.

3.3.3 Reliable and Secure Edge AI Deployments:

- Ensuring AI model integrity using secure hardware accelerators and encrypted model updates.
- Deploying AI-powered cybersecurity applications for threat detection in Edge Computing networks.

3.3.4 Seamless Cloud-Edge AI Orchestration:

- Enabling collaboration between cloud and edge AI models for intelligent workload distribution.
- Developing adaptive AI pipelines that dynamically adjust computational loads based on network health.

The Edge Intelligence research roadmap indicates the architectural layers, optimization objectives, and open issues in combining AI and Edge Computing. Through performance, cost, privacy, efficiency, and reliability challenges, innovations in the future will fuel the next AI-fueled Edge Computing wave, transforming smart cities, autonomous systems, industrial automation, and many more.

3.4 Recapitulation of Intelligence-Enabled Edge Computing (IEC)

The left of the Edge Intelligence research roadmap depicts AI for Edge, or what we refer to as Intelligence-Enabled Edge Computing (IEC). IEC utilizes AI methods to improve Edge Computing, avoiding complicated decision-making, learning, and resource management challenges. Through the integration of AI, Edge Computing is more efficient, self-sustaining, and adaptive and addresses key problems in system orchestration, information management, and real-time service delivery. From a bottom-up viewpoint, IEC researches may be segregated into three highest levels: Topology, Content, and Service. In every level, there are corresponding problems that may be solved through AI, raising the overall productivity of Edge Computing systems.

The Topology layer is concerned with the structure, deployment, and management of Edge Computing devices in the most productive way and at the minimum operational expense. Orchestration of Edge Sites (OES) and Wireless Networking (WN) are two significant topics in topology optimization. OES deals with deployment and management of edge sites, which are small data centers that are plugged into Small-Cell Base Stations (SBSs). Artificial intelligence-based orchestration guarantees optimum deployment and management of edge resources, which guarantees optimum delivery of services. One emerging edge orchestration research direction is mobile edge servers using Unmanned Aerial Vehicles (UAVs). Artificial Intelligence-based access points on UAVs can dynamically adjust to network conditions, optimizing coverage, energy efficiency, and service availability. AI-based scheduling and flight planning enable UAVs to predict user mobility and pre-fetch data, enhancing Quality of Experience (QoE).

Wireless Networking (WN) in Edge Computing includes data collection and network planning, both of which are enhanced by AI. Data collection is concerned with quick and efficient collection of dispersed data from IoT devices, while network planning entails radio resource management, access control, and network scheduling. AI models enhance wireless network performance with congestion prediction, bandwidth adjustment of allocation, and enhanced signal encoding and decoding. Machine learning algorithms are increasingly being incorporated into smart networking systems, constructing self-optimizing wireless communication procedures. Deep Reinforcement Learning (DRL) methods, like Deep Q-Networks (DQNs), have been used for adaptive radio resource control, data transmission optimization, and interference cancellation in Fog Radio Access Networks (F-RANs).

The Content layer in IEC addresses data provisioning, service placement, service composition, and caching. AI methods are employed for improving delivery of services and management of most accessed information. Data and service provisioning assures the proper distribution of computational power between cloud servers, edge nodes, and cellular devices. AI-based QoS-aware systems have been developed to direct data in an enlightened manner for seamless delivery of services. Service placement specifies where and how the services are allocated at the edge, and AI-based service placement models can dynamically adjust resources based on demand changes, infrastructure constraints, and

user mobility patterns. RL models such as multi-armed bandit algorithms have been utilized heavily for optimal resource utilization under the budget.

Service composition is AI-based service selection for energy minimization and user QoE maximization. AI-driven service composition platforms are being designed that choose the optimal service components automatically so that edge applications execute with negligible latency and power consumption. Service caching also becomes an essential component of Edge Computing by caching highly accessed data and services at key edge points. AI-driven caching platforms can forecast most-accessed data and thus enable pre-emptive caching of content and group-level cache management across numerous edge devices. Multi-agent learning allows service caching platforms to optimize performance in large-scale edge systems by minimizing data transfer latency and network usage.

The IEC Service layer also solves three major challenges: computation offloading, user profile migration, and mobility management. Computation offloading allows offloading of processing workloads from edge devices to remote cloud servers, so computations are performed where they can do it best. AI-based decision-making models decide whether a computation task is to be executed locally or offloaded, depending on the network condition, device ability, and processing priority. Lyapunov optimization techniques and Deep Q-Networks (DQNs) are employed to optimize multi-server, multi-user computation offloading strategies on a large scale. AI models reduce latency, energy usage, and bandwidth to ensure maximum usage of resources in edge networks.

User profile migration sidesteps the issue of user settings, logs, and customized settings migration as users move across various edge networks. Migration techniques based on artificial intelligence provide continual service with intelligent migration adaptation responding in real-time to variations in user location, network, and application requirements. Future AI techniques are emerging to create cooperative migration models that help streamline profile adaptation through predictive analytics and prior usage information. Mobility management, the second key building block of the Service layer, leverages AI to forecast and optimize user mobility patterns. AI-based mobility models facilitate anticipatory service provisioning so that the users have consistent network quality and seamless availability of service while moving from one edge zone to another. Machine learning algorithms are applied more and more to dynamically share network resources, optimally handover, and enhance service continuity in mobile edge environments.

In short, AI is of paramount importance to the development of Intelligence-Enabled Edge Computing (IEC) in that it enables intelligent resource allocation, predictive service placement, dynamic cache policies, and real-time decision-making capabilities. AI-based optimizations enable Edge Computing systems to adapt to varying workloads, lower operational expenses, and improve efficiency overall. Upcoming IEC research will be geared towards creating ultra-efficient AI models for low-power edge environments, decentralized AI training to minimize cloud reliance, more advanced AI-based security features, and designing autonomous self-optimizing Edge Computing frameworks. As Edge Computing and AI keep progressing, IEC will be the backbone for all future smart systems, supporting real-time automation, effortless integration of AI, and distributed intelligence in a broad scope of applications ranging from smart cities and autonomous cars to industrial automation and 5G networks.

3.5 Recapitulation of Artificial Intelligence on Edge (AIE)

Right edge of the Edge Intelligence research roadmap represents AI on Edge (AIE) and focuses on AI model training and inference to be performed directly at the edge of the network. Whereas AI is utilized for Edge Computing optimization in Intelligence-Enabled Edge Computing (IEC), AIE considers the role of deploying AI models in a efficient way on edge servers and devices, and training them and executing them. According to the top-down framework, AIE research constitutes three broad areas: Framework Design, Model Adaptation, and Processor Acceleration. Together, these areas catalog the ways that AI is combined into Edge Computing ecosystems to render edge AI processing with privacy preserving, low-latency, and decentralization.

Framework Design targets building useful architecture for the training and inference of edge AI models. Rather than developing modifications to existing AI models, researchers focus on creating frameworks that improve distributed model training and inference without the need for significant updates in deep learning models. With the exception of knowledge distillation-based Model Training frameworks, all suggested Model Training frameworks are distributed. Data splitting and model splitting are forms of distributed training methodologies. Data splitting is responsible for splitting training sets between edge devices and whether or not models are combined through master-device, helper-device, or device-device coordination. Model splitting splits neural network layers between distinct edge nodes, balancing computational performance through collaborative learning across multiple devices. Knowledge distillation-based systems improve model efficiency by transferring knowledge from large, complicated AI models to small,

lightweight models that can be executed effectively on resource-limited edge devices. The method enables AI models to be trained in cloud and then fine-tuned at the edge with little computation overhead but high accuracy.

Federated Learning is one of the most popular frameworks for AI model training at the edge. In contrast to the conventional methods of AI training on the concentration of large volumes of data in a central cloud, Federated Learning allows local model training on various edge devices with only model updates being exchanged and not raw data. This approach greatly improves privacy and security, especially for applications such as healthcare, finance, and other sensitive AI uses. But Federated Learning brings new problems first and most importantly of these is communication efficiency because entire local models need to be updated regularly and sent to a central server to be summed. Researchers are striving to cut down communication overhead and optimize model aggregation methods for making Federated Learning at the edge more powerful.

In device-based distributed AI training, Stochastic Gradient Descent (SGD) continues to be an elementary optimization method. Edge devices utilize SGD to locally update the model parameters from their own datasets, considering each update as a mini-batch. Local updates are forwarded to a central node to aggregate the global model. The key challenge of such an approach is how to trade off between model performance and communication efficiency. If all the local gradients of edge devices are sent simultaneously, it may result in network congestion and higher latency. Another effective strategy is selective gradient sharing, wherein only model updates with considerable gains are shared to avoid communication overhead while not compromising model accuracy.

For Model Inference, partitioning of the model is possibly the most prevalent paradigm. According to this paradigm, deep learning models are partitioned into multiple components, compute-intensive layers are computed on edge servers with performance as a focus, and light layers are executed on mobile or IoT devices directly. The model partitioning problem is how to select the best split point, ensuring inference accuracy to be as close to original accuracy as possible while keeping the computational load on low-capability edge devices low. Model compression, input filtering, and early-exit are some other model inference methods that minimize latency and energy consumption to make AI inference viable on edge hardware.

Model Adaptation is another important area of AI on Edge that addresses adapting current AI training and inference architectures to be deployable in the edge. Most prevalent deep learning models are cloud-optimized since computation is practically unlimited in clouds. In Edge Computing, however, AI models must be optimized for deployment in low-power decentralized settings. There are a number of approaches that are adopted to convert AI models for the edge, i.e., Model Compression, Conditional Computation, Algorithm Asynchronization, and Thorough Decentralization.

Model Compression decreases AI model sizes by eliminating redundant parameters without compromising performance. Quantization, dimensionality reduction, pruning, and precision downgrading are a few methods utilized to optimize resource-limited edge environments for models. Some of the most widely used techniques applied in model compression are Singular Value Decomposition (SVD), Huffman Coding, and Principal Component Analysis (PCA). Conditional Computation also reduces computation load by selectively turning off computations not required in deep learning models dynamically, through techniques such as components shutoff, input filtering, early exit, and results caching. These techniques guarantee efficient AI inference on low-processing capacity devices.

Algorithm Asynchronization is another method of adaptation for optimizing model aggregation in Federated Learning and other distributed AI frameworks. Rather than waiting for all the edge devices to finish their local updates before synchronizing models, asynchronous model updates enable more efficient edge devices to update at more regular intervals, thereby enhancing training efficiency. Full Decentralization entirely eliminates the requirement for a centralized model aggregator, thus eliminating the problem of server failure and exposure to privacy attacks. Technologies like blockchain-based decentralized learning platforms and game-theoretic AI optimization methods are being researched to remove central dependencies without compromising model integrity.

Processor Acceleration addresses hardware optimizations that accelerate the performance of deep learning models at the edge. As AI inference and training are MAC-intensive, efficient hardware acceleration is required to boost performance. Scientists are designing special instruction sets for deep learning computation, massively parallelized computing models, and near-data processing hardware platforms. Massively parallel computing models come under temporal and spatial models. Temporal models, including CPUs and GPUs, obtain acceleration by minimizing redundant multiplications and maximizing the processing throughput. Spatial architectures, for example, Field Programmable Gate Arrays (FPGAs), Tensor Processing Units (TPUs), and Neural Processing Units (NPUs), have optimal data reuse and efficiency through hardware optimization for AI workloads.

Near-data processing is also a promising direction that brings computation near memory to mitigate data transfer latency and enhance energy efficiency. AI hardware co-design is increasingly becoming a critical part of Edge Computing, making sure that AI models are well optimized for the host hardware architecture. Over the last few years, researchers have been concentrating on integrating memristor crossbar arrays and adaptive processing units to further speed up deep learning inference at the edge. However, since processor acceleration is predominantly studied by AI hardware researchers, the paper does not dive into the topic in detail, and readers may look at specialist literature on deep learning hardware acceleration.

In short, Artificial Intelligence on Edge (AIE) is a paradigm shift in the deployment of AI to enable real-time, distributed intelligence in an enormous variety of applications. With better training paradigms, better model adaptation methods, and leveraging hardware accelerations, AIE provides a guarantee that AI models operate optimally within resource-scarce environments. The next field of AIE research will be the creation of ultra-lightweight AI models for edge devices, enhancing Federated Learning efficiency, building AI security at the edge, and further advancing AI chip architectures that are specifically designed for decentralized deep learning. As the market for low-latency, privacy-preserving AI solutions expands, AIE will be at the vanguard of defining the next generation of smart edge systems, from smart cities to autonomous cars and beyond.

4 AI on Edge

Within Subsection III-C, we sorted activities in AI on Edge (AIE) research into three general categories: Model Adaptation, Framework Design, and Processor Acceleration. Training and inference framework design of AI on edge is still limited to that of cloud-based AI models, mainly because of limitations in compute, storage, and networking bandwidth. Training paradigms like Federated Learning and Knowledge Distillation are intended to enable distributed AI model training on the edge, while inference paradigms like Model Splitting and Model Partitioning are intended to optimize AI computation on computationally limited edge devices. Since the edge device has much less computational power than cloud servers, a super research challenge is making AI models that are efficient to train and run on the limited edge resources.

Rather than designing new training and inference AI frameworks from scratch, most of the effort goes into optimizing already existing frameworks in an attempt to edge-deploy them more cost-effectively. A lot of work in AI for Edge is Model Adaptation, with optimization of Federated Learning to minimize resource usage while retaining model accuracy. Processor Acceleration entails low-level hardware optimizations, so no discussion of it is provided here. The subsequent sections explore Model Adaptation further, highlighting the most recent advances in research, principal techniques, and grand challenges in taking AI to the edge.

4.1 State of the Art in AI on Edge

4.1.1 Model Compression: Optimizing AI Models for Edge Execution

Model Compression methods try to minimize the computational, memory, and bandwidth usage of AI models by exploiting the intrinsic sparsity of weights and gradients. Quantization, Dimensionality Reduction, Pruning, Precision Downgrading, and Component Sharing are some of the prominent methods. These methods help keep AI models light and thin with inference accuracy intact.

Some sophisticated methods have been researched to compress AI models with accuracy retained. For example, certain work suggests sketched and organized updates to reduce the communication overhead of Federated Learning by compressing the model updates before sending. Secure aggregation that attempts to balance communication efficiency with privacy is another solution suggested. Other work suggests that Deep Neural Networks (DNNs) tend to be over-parameterized, i.e., have redundant weights that can be pruned without taking a heavy toll on accuracy. A retrain-after-pruning strategy is proposed by some researchers, where a DNN is pruned and then retrained on new data such that performance is made stable without losing model complexity.

Quantization methods like Binary Neural Networks (BNNs) and low-bitwidth mixed precision compression substitute lower-bit or binary representations for full-precision model weights, reducing memory by orders of magnitude but increasing inference. Hybrid models with binary and full-precision layers also improve energy efficiency without performance loss. Hybrid Matrix Decomposition (HMD) is a new method that introduces a partitioned matrix structure that speeds up inference by concentrating computational resources on the most pertinent sections of a model.

Partitioned DNN-based compression is another trend, in which AI models are divided between edge devices and cloud servers. In some methods, the early layers of a model are quantized for running on edge devices with deeper layers in full precision on the cloud. This offloads the workload from the edge while preserving high accuracy. Another approach is to create classifiers at the mid-model layer levels so that AI inference can halt prematurely when confidence levels are attained, reducing overall computation time.

4.1.2 Conditional Computation: Dynamic AI Execution for Efficiency

Conditional Computation methods block unnecessary computation in deep learning models selectively, enhancing efficiency without affecting accuracy. AI models can be dynamically tuned at runtime to reduce resource consumption by using methods like Component Shutoff, Input Filtering, Early Exit, and Results Caching.

Another popular use of Conditional Computation is block-wise dropout, in which AI models are dynamically throttled to acquire correct results using less resources. Runtime-adaptive neural networks shut down certain portions of the model during inference time to minimize computation overhead. The concept has also been applied to Federated Learning, in which only the most precious edge devices contribute to training and prune the less precious contributors to reduce model aggregation complexity and minimize communication latency.

4.1.3 Algorithm Asynchronization: Edge AI Model Training Acceleration

Algorithm Asynchronization methods are implemented to minimize communication bottlenecks within Federated Learning by supporting asynchronous AI model aggregation. On edge networks, where connectivity is not consistent and device availability cannot be counted upon, synchronous training methods can be less than ideal.

One such solution is Random-Gossip Communication, in which model updates are shared peer-to-peer instead of via a central server. The method minimizes the danger of network saturation and enhances convergence efficiency as a whole. Explore some research recommends GoSGD, an asynchronous stochastic gradient descent algorithm for distributed-scale AI training. Toward further enhancement in convergence rates, GossipGraD recommends a communication-efficient variant of gossip-based AI training that minimizes overhead while preserving accuracy.

4.1.4 Total Decentralization: Privacy-Preserving AI on the Edge

Total Decentralization abolishes a central model aggregator to minimize the chance of privacy breaches in Federated Learning. Although Federated Learning keeps training data resident locally on edge devices, model updates still contain sensitive information that may be hacked.

Game-theoretic AI optimization methods and blockchain technology provide solutions of huge promise for complete decentralized AI training. Blockchain-based Federated Learning systems like BlockFL utilize smart contracts to relieve themselves of central-server reliance. Here, edge devices serve as miners who exchange model updates securely and authenticate contributions according to Proof-of-Work (PoW) or alternative consensus algorithms. Decentralized architecture here increases security and allows AI models to be co-trained without needing the trust of one party.

4.2 Grand Challenges in AI on Edge

There are still grand challenges with AI on Edge in data availability, model choice, and coordination mechanisms.

4.2.1 Data Availability: Having Useful and Varied Training Data

Availability and quality of data are the biggest issue in AI on Edge. The models of AI need large, diverse, and well-spread data for maximum accuracy. The user data is biased, noisy, and non-i.i.d., which makes it harder to train stable models.

To address this, incentive models could be made mandatory to encourage data sharing between edge users. Additionally, new robust AI training procedures need to be designed to efficiently handle imbalanced and non-uniform datasets.

4.2.2 Model Selection: Optimizing AI Models for Edge Environments

Choosing the right configurations of AI models, training configurations, and hardware accelerators is a problem. The models should be optimized to be efficient for immediate deployment and utilization of resources and incorporate selective choice of learning boundaries, compression methods, and hardware parameters. As model selection is related to resource allocation and device capacity constraints, it is an open problem to design self-adaptive AI model adaptation strategies for heterogeneous edges.

4.2.3 Coordination Mechanisms: Coordination of Heterogeneous Edge Devices

Edge AI has to execute on highly heterogeneous hardware and software platforms, such as IoT devices, smartphones, embedded systems, and dedicated AI accelerators. The compatibility and coordination of these devices have to be ensured for the deployment of scalable AI.

Efforts in the future will need to center around building coordinated middleware stacks and APIs that support frictionless execution of AI on various edge platforms. A coordination system would enable AI models to dynamically change their computation plan based on runtime network conditions and device capabilities.

5 Conclusion

Edge Intelligence, in its nascent development stage, has attracted a lot of attention from industries and researchers. Its promise to transform data processing and AI applications at the network edge has been provoking more research efforts and technological innovations. This paper seeks to present a structured and thorough summary of research opportunities and future challenges in Edge Intelligence by presenting a well-defined classification framework.

To build a firm foundation, we initially discussed how Artificial Intelligence and Edge Computing complement and augment each other and explored the relationship between these two domains. AI gives Edge Computing sophisticated optimization, learning, and decision-making capability, allowing dynamic resource allocation, optimal data processing, and smart delivery of services. Edge Computing, in return, brings the promise of AI closer with real-world scenarios of deployment, decentralized learning systems, and low-latency support for AI-driven services. The symbiotic combination of Edge Computing and AI is the cornerstone of Edge Intelligence, fueling innovation in various application fields including smart cities, autonomous driving, and industrial control.

- To further delineate the research space, we divided Edge Intelligence into two broad categories: AI on Edge and AI for Edge, and developed a well-structured research roadmap for each.

- AI for Edge (Intelligence-Enabled Edge Computing - IEC) investigates the potential of AI in optimizing Edge Computing by solving the difficult learning, planning, and resource allocation problems.

Using a bottom-up design, we organized research activity in Topology, Content, and Service, proving how AI can be used to improve network orchestration, data provisioning, service placement, and mobility management. Artificial intelligence-based methods like reinforcement learning, federated learning, and deep neural networks have been used for intelligent networking, real-time data caching, and service provisioning that provide scalable, efficient, and adaptive solutions in edge scenarios. •\tAI on Edge (Artificial Intelligence on Edge - AIE) examines how AI models should be effectively trained and run on constrained edge devices. Through top-down decomposition, we broke down research into Model Adaptation, Framework Design, and Processor Acceleration. We reviewed current frameworks for distributed AI training and inference, such as Federated Learning, Model Splitting, and Knowledge Distillation, and explored optimization techniques like model compression, quantization, algorithm asynchronization, and decentralized AI coordination.

Furthermore, we underlined today's state of the art and provided important research challenges in AI for Edge and AI on Edge. Among them, some of the most significant ones are data availability and bias, best model choice, unobtrusive coordination among heterogeneous edge devices, and decentralized AI security. The solving of these issues will be critical to continuing progress in scalable, privacy-protecting, and high-performance AI solutions at the edge.

With ongoing progress in Edge Intelligence, it is likely to revolutionize existing computing paradigms through the enablement of low-latency AI inference, real-time decision-making, and autonomous learning across distributed networks. Future research will have to encompass the design of energy-efficient AI models, the design of Federated Learning frameworks, the development of AI-driven security controls, and edge hardware design optimization.

References

[1] G. P. A. W. Group, “View on 5G architecture: Version 3.0,” June 2019.

[2] M. Asif-Ur-Rahman, F. Afsana, M. Mahmud, M. S. Kaiser, M. R. Ahmed, O. Kaiwartya, and A. James-Taylor, “Toward a heterogeneous mist, fog, and cloud-based framework for the internet of healthcare things,” IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4049–4062, June 2019.

[3] Ericsson, "IoT connections outlook: NB-IoT and Cat-M technologies will account for close to 45 percent of cellular IoT connections in 2024," Ericsson Mobility Report, June 2019.

[4] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," IEEE Internet of Things Journal, vol. 5, no. 1, pp. 450–465, Feb 2018.

[5] ETSI, "Multi-access edge computing (MEC): Study on MEC support for alternative virtualization technologies," November 2019.

[6] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," 2018.

[7] Q. Mao, F. Hu, and Q. Hao, "Deep learning for intelligent wireless networks: A comprehensive survey," IEEE Communications Surveys Tutorials, vol. 20, no. 4, pp. 2595–2621, Fourth quarter 2018.

[8] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Machine learning for wireless networks with artificial intelligence: A tutorial on neural networks," 2017.

[9] M. Peng and K. Zhang, "Recent advances in fog radio access networks: Performance analysis and radio resource allocation," IEEE Access, vol. 4, pp. 5003–5009, 2016.

[10] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," Proceedings of the IEEE, vol. 107, no. 8, pp. 1738–1762, Aug 2019.

[11] X. Zhang, Y. Wang, S. Lu, L. Liu, L. Xu, and W. Shi, "OpenEI: An open framework for edge intelligence," 2019.

[12] Y. Han, X. Wang, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," 2019.

[13] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," IEEE Network, vol. 30, no. 4, pp. 46–53, July 2016.

[14] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications," IEEE Internet of Things Journal, vol. 4, no. 5, pp. 1125–1142, Oct 2017.

[15] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," AISTATS, 2016.

[16] J. Xu, Y. Zeng, and R. Zhang, "UAV-enabled wireless power transfer: Trajectory design and energy optimization," IEEE Transactions on Wireless Communications, vol. 17, no. 8, pp. 5092–5106, Aug 2018.

[17] B. Li, Z. Fei, and Y. Zhang, "UAV communications for 5G and beyond: Recent advances and future trends," IEEE Internet of Things Journal, vol. 6, no. 2, pp. 2241–2263, April 2019.

[18] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," IEEE Journal on Selected Areas in Communications, vol. 35, no. 5, pp. 1046–1061, May 2017.

[19] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Towards an intelligent edge: Wireless communication meets machine learning," 2018.

[20] Y. Sun, M. Peng, and S. Mao, "Deep reinforcement learning-based mode selection and resource management for green fog radio access networks," IEEE Internet of Things Journal, vol. 6, no. 2, pp. 1960–1971, April 2019.