Paper ID : J240478

AI-Empowered IoV Architecture for Real-time Application in heterogeneous 6G

Linyu Zhu[1] ; Song Bao[2] ; Shoupeng Lu[2]

[1] College of Software Engineering, Sichuan University, Chengdu, China
[2] College of Computer Science, Sichuan University, Chengdu, China

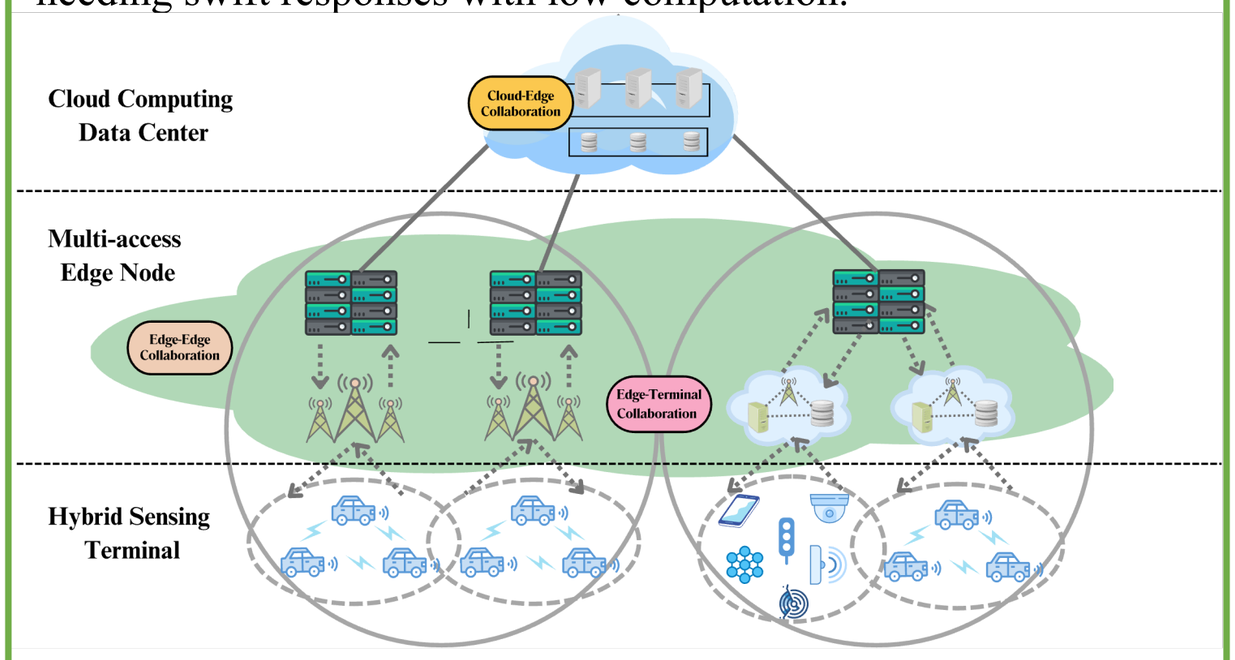
Introduction

The convergence of AI, IoV, 6G, and sensors is propelling traditional connected car services towards intelligent connectivity. AIoV facilitates data exchange, leveraging wireless tech for road traffic safety. However, challenges like transmission latency and high-speed data processing require innovative solutions. Traditional cloud computing falls short, prompting the proposed Cloud-Edge-Terminal AI architecture. This distributes processing for real-time efficiency, reducing latency and enhancing response speed. This article introduces the architecture, key technologies, application scenarios, and future directions of AIoV. Sections include architecture proposal, application scenarios, and future directions.

Cloud-Edge-Terminal Collaboration Architecture and Technology

Architecture

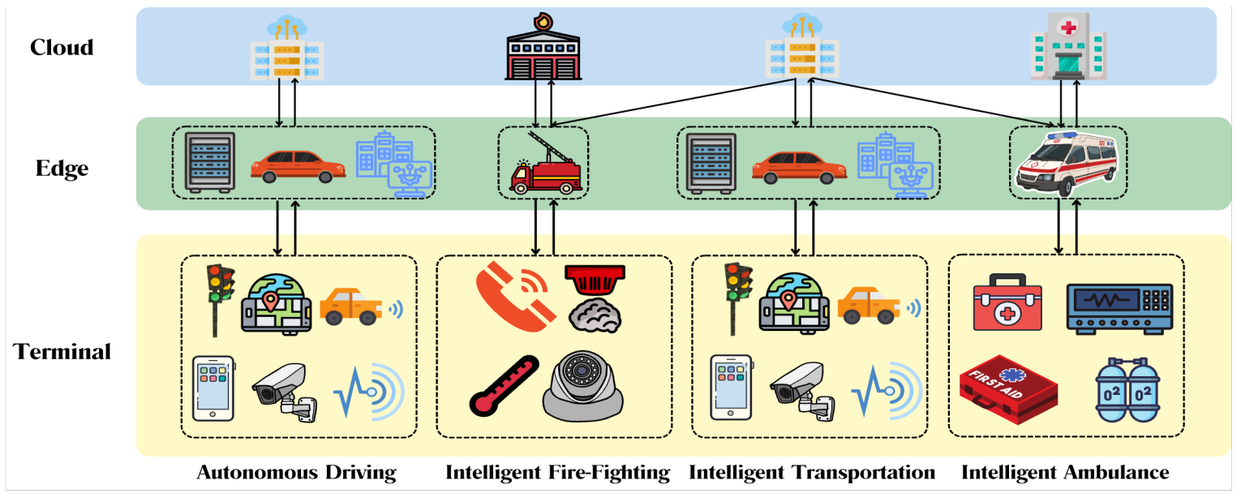
Cloud-Oriented Cloud-Edge Collaboration: Clouds handle model training and update edge models, reducing edge computational load.
Edge-Oriented Cloud-Edge Collaboration: Clouds train generic models and transmit them to edges, which fine-tune locally.
Dual-Oriented Cloud-Edge Collaboration: Combines cloud and edge efforts for bidirectional model optimization. **Vertical Edge-Edge Collaboration**: Decentralized task execution among edges for load balancing and optimal resource use. Intelligent task assignment and scheduling enhance processing speed.
Horizontal Edge-Edge Collaboration: Features a coordinator edge for task distribution and a server edge for task execution, focusing on scalability and flexible resource management.
Hybrid Edge-Edge Collaboration: Integrates vertical and horizontal approaches for comprehensive edge cooperation.
Edge-Terminal Collaboration: Terminals collect data for edge processing and receive instructions for action. Ideal for tasks needing swift responses with low computation.



Technology

Task Offloading: This process involves transferring tasks to locations with richer resources to optimize utilization and reduce latency. Strategies must balance scheduling and migration costs, ensuring real-time edge response with cloud computing power. Research trends include mathematical optimization and deep learning approaches.
Model Compression: A critical step in offloading, model compression reduces model size for efficient communication without significantly sacrificing quality. Techniques include frontend (e.g., knowledge distillation, network design) and backend (e.g., low-rank approximation, quantization) methods, with the latter potentially altering the network structure.
Federated Learning: FL facilitates distributed model training across edge clients, using local data to update models and aggregating them into a global model at a central server. Aggregation strategies are either synchronous or asynchronous, with current research exploring hybrid models for efficiency and performance.
Transfer Learning: Addressing virtual concept drift in IoV, transfer learning methods (model-based, data-based, and query-based) aim to enhance cross-domain adaptability and generalization. The focus is on fast knowledge transfer, deep knowledge transfer, and privacy security.
Edge Caching: Edge servers cache content to minimize retrieval latency and network congestion. Due to limited capacity, cache replacement schemes are crucial for improving hit rates and service latency. Notably, reinforcement learning-based strategies are effective in dynamic network environments.

AIoV Use Cases and Application Scenarios



Future Directions and Works

Tensor-based Multimodal Data Compression

- Tensors effectively represent complex data, aiding in deep compression and acceleration on edge devices.
- Balancing compression and accuracy is a key optimization challenge.

Large and Small Models Collaboration

- A computational collaborative model is sought to combine cloud and edge computing for practical applications.
- Future research will focus on high-precision model compression and cloud-edge-device co-evolution.

Hardware-Software Integrated Edge Intelligence Acceleration

- Hardware design should consider algorithm demands and support real-time processing.
- Algorithm development should be aligned with hardware capabilities for efficiency and low-latency operations.

Conclusion

This paper envisions AIoV enhancing system performance through a Cloud-Edge-Terminal architecture, addressing data processing and energy constraints. It introduces key collaboration technologies for advanced inference tasks and sustainable services, presents practical applications for daily convenience, and identifies challenges and opportunities for further research in AIoV collaborative systems.