Edge AI for Electric Vehicles: Smarter Models at the Edge

As electric vehicles become more prevalent, the need for intelligent onboard systems to monitor battery health and predict faults becomes increasingly critical. But while deep neural networks have shown remarkable capabilities in these tasks, deploying such models on the low-powered microcontrollers found inside vehicles presents a significant challenge. These embedded devices often lack the computational power, energy capacity, and connectivity required to run complex AI models. At the same time, privacy concerns prevent sensitive data from being sent to centralized servers for processing.

In this work, we present a novel architecture designed to bring AI directly to the edge, making it viable even in resource-constrained environments like those found in electric vehicles. By integrating Federated Learning, Knowledge Distillation, and model compression, we enable what we define as Edge Intelligence—an approach that moves both the training and inference

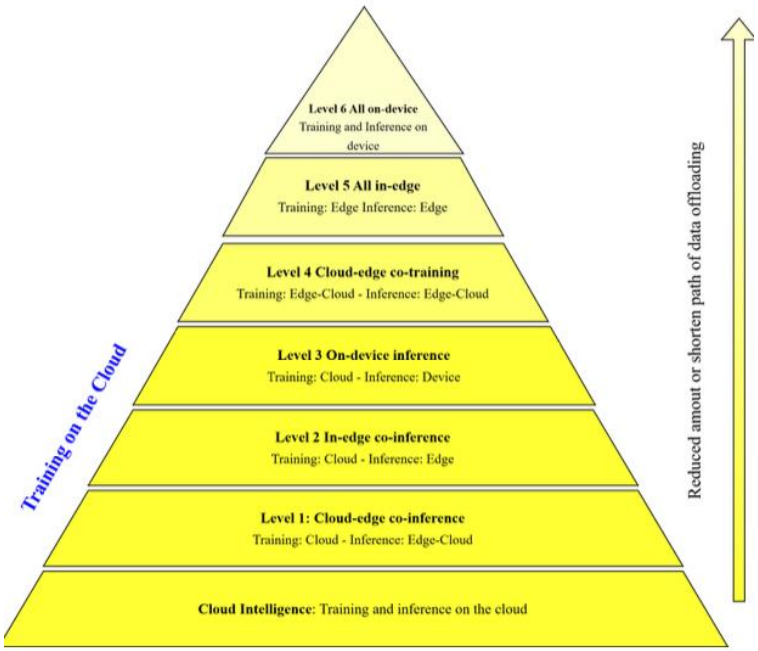

Figure 3.1 Six-level rating for EI described in [9].

of AI models closer to the data source, improving efficiency and preserving privacy.

In this context, edge devices in electric vehicles are not just passive sensors but become active participants in learning. Rather than sending raw data to the cloud, each vehicle locally trains a machine learning model using its own data and then sends only the updated model weights

to a central server. This process, known as Federated Learning, preserves privacy and reduces bandwidth usage while still allowing for a shared, continually improving model to emerge. The cloud server acts as a coordinator, aggregating model updates from multiple

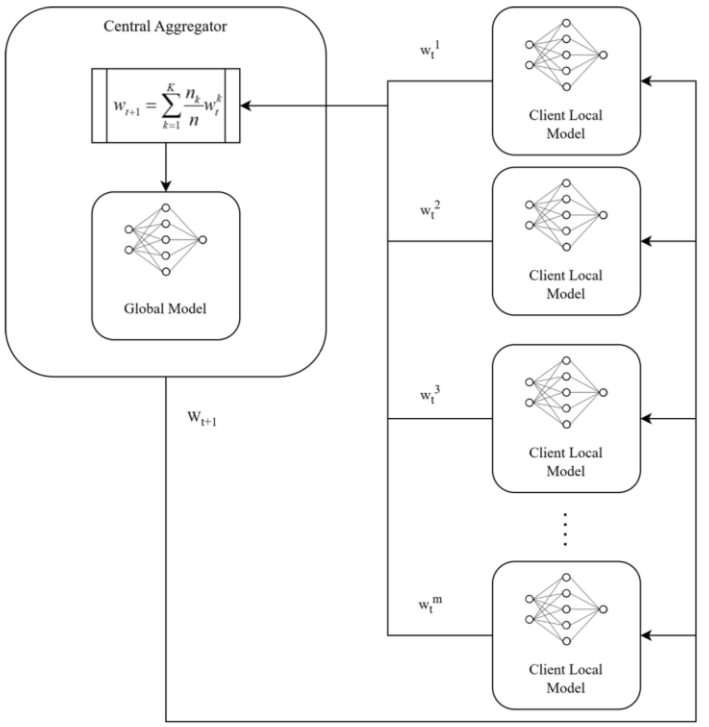

Figure 3.2 In a Federated Learning scenario, each client trains its model leveraging its own private data and sends its model parameters to a central server. The central server aggregates the parameters received from each client to enhance the performance of the central global model, which is then sent back to the clients.

vehicles and redistributing the refined model back to the fleet [6].

However, Federated Learning alone is not enough. The AI models used in real-world applications are often too large and computationally intensive to run on microcontrollers. That's where Knowledge Distillation comes into play. In this approach, a large, complex "teacher" model trained in the cloud transfers its knowledge to a smaller "student" model that

mimics its behavior [3]. The student model is much lighter and faster, making it suitable for deployment on embedded devices without significant sacrifices in performance.

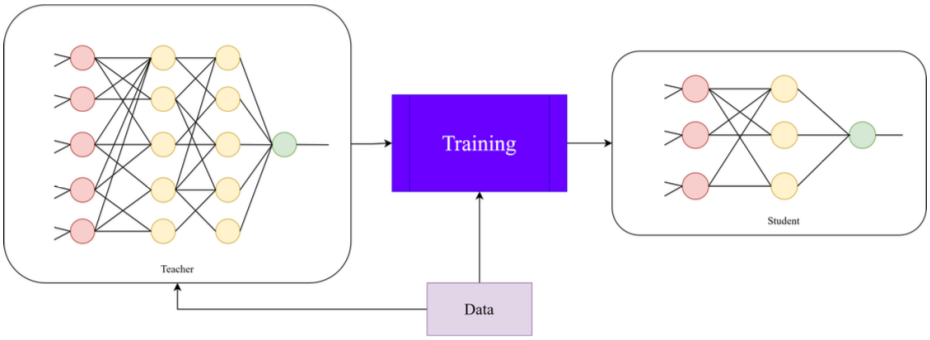

Figure 3.3 The schema illustrates the fundamental concept of KD: during the training of a simplified neural network, knowledge from a larger network is transferred to the smaller one.

To further optimize these student models, compression techniques such as quantization and pruning are applied. Quantization reduces the precision of the model's weights, leading to smaller models and faster inference [1]. Pruning eliminates unnecessary parameters to streamline the model architecture [2]. These techniques allow even relatively complex neural networks to be reduced in size and adapted to the constraints of edge hardware, all without major losses in accuracy [5].

One of the most compelling use cases explored in the proposed architecture is the monitoring of lithium-ion batteries in electric vehicles. These systems must be able to estimate key parameters like the state of charge or state of health in real time, using data generated from within the vehicle. By executing the AI models directly on the vehicle, real-time responses are possible even when a stable internet connection is not. This is particularly important for

features like predictive maintenance, which relies on up-to-date insights into battery

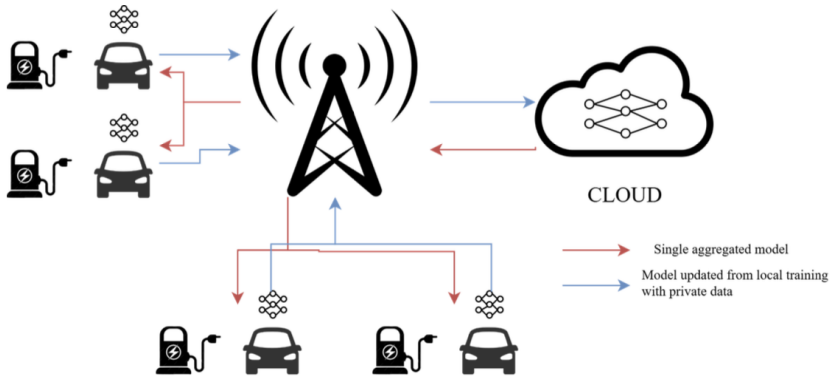

Figure 3.4 Use case scenario.

behavior to prevent failures [4].

The entire workflow begins in the cloud, where a complex neural network is trained using centralized data. This model is then distilled, compressed, and exported in formats like ONNX or TinyML, compatible with microcontroller-based platforms. Through over-the-air updates, the optimized models are distributed to vehicles, where they perform real-time inference and continue training locally. When a vehicle reconnects to the network, it can share the updated model weights back to the cloud, contributing to the collective learning process without ever transmitting sensitive raw data [8].

The architecture designed by the researchers includes a central Cluster Aggregator deployed in the cloud and a network of Distributed Agents residing on edge devices in each vehicle. The Cluster Aggregator manages the overall federated learning process, handles model aggregation, monitors the state of participating clients, and performs training and distillation tasks [7]. It also includes modules for model compression and secure deployment. On the vehicle side, each Distributed Agent handles local inference and training, synchronizes with

the cloud when connectivity is available, and ensures that the vehicle can function autonomously when it's not.

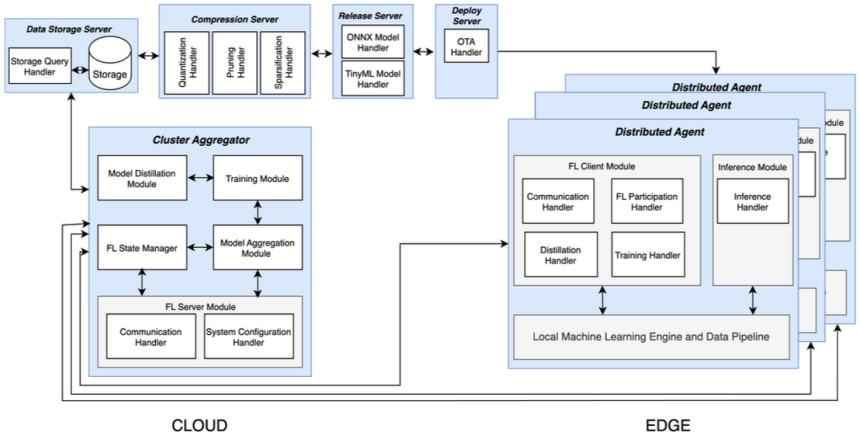

Figure 3.7 The final architecture includes a Cluster Aggregator, deployed in the Cloud, and Distributed Agents, deployed on resource-constrained edge devices.

Designing such a system comes with challenges. Vehicles are mobile, and their internet connections are not always reliable. The hardware they use varies widely, and ensuring compatibility across devices with different processing capabilities can be complex. There's also a delicate balance to strike between training accuracy and the size and speed of the models—too much compression can hurt performance, while too little can overwhelm the device.

Despite these challenges, the proposed architecture demonstrates that it is possible to bring high-performance AI to the edge, even in demanding environments like electric vehicles. By combining federated learning with knowledge distillation and efficient model compression, the system achieves a powerful trifecta: privacy, performance, and portability. This is not just a theoretical advancement but a practical roadmap for building the next generation of intelligent, self-improving vehicles.

This article illustrates one of the key research areas explored in the European project NEUROKIT2E, which aims to develop open and efficient Edge AI technologies for embedded systems, including applications in electric mobility.

References

1. *Y. Cai et al., "ZeroQ: A Novel Zero Shot Quantization Framework," CVPR, 2020.*
2. *Hanson, S. J. & Pratt, L. Y. "Comparing biases for minimal network construction," NIPS, 1988.*
3. *Hinton, G.E., Vinyals, O., & Dean, J. "Distilling the Knowledge in a Neural Network," arXiv:1503.02531.*
4. *Yi Li et al., "Data-driven health estimation of lithium-ion batteries," Renewable and Sustainable Energy Reviews, 2019.*
5. *Li, Z.; Li, H.; Meng, L. "Model Compression for Deep Neural Networks: A Survey," Computers, 2023.*
6. *McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," AISTATS, 2016.*
7. *Mora et al., "Knowledge Distillation in Federated Learning: A Practical Guide," IJCAI-24, 2024.*
8. *Polino et al., "Model Compression via Distillation and Quantization," arXiv:1802.05668.*
9. *Zhou et al., "Edge Intelligence: Paving the Last Mile of AI," Proceedings of the IEEE, 2019.*