Cross-Modal Knowledge Distillation for Ultra-Lightweight Edge AI: A Resource-Aware Approach

Aaron Mano Cherian | Dr. Manoov R | School of Computer Science and Engineering

Introduction

Edge computing applications face significant challenges when deploying complex multi-modal AI models due to severe resource constraints. Current approaches sacrifice performance for efficiency, creating a critical need for solutions that maintain accuracy while dramatically reducing resource requirements. This project introduces a Resource-Aware Cross-Modal Knowledge Distillation (RA-CMKD) framework to bridge this gap and enable sophisticated AI capabilities on resource-limited edge devices. The framework implements three strategic innovations: a dynamic resource allocation module, modality-specific compression techniques, and lightweight feature alignment strategies, all designed to work within a three-tier hierarchical system architecture spanning cloud, fog, and edge components.

Motivation

The deployment of multi-modal deep learning models on edge devices presents several critical challenges that necessitate innovative solutions. Edge devices inherently operate with restricted processing power, memory capacity, and energy resources, making the deployment of sophisticated AI models particularly difficult. Existing model optimization approaches often result in significant performance degradation, limiting their practical utility. Furthermore, current solutions lack dynamic adaptation capabilities, making them unsuitable for the fluctuating resource conditions typical in edge environments. Knowledge transfer between different modalities becomes especially challenging in resource-constrained settings, compounding the complexity of multi-modal applications. Additionally, the deployment across numerous heterogeneous devices introduces significant maintenance challenges that must be addressed for practical implementation.

Scope of the Project

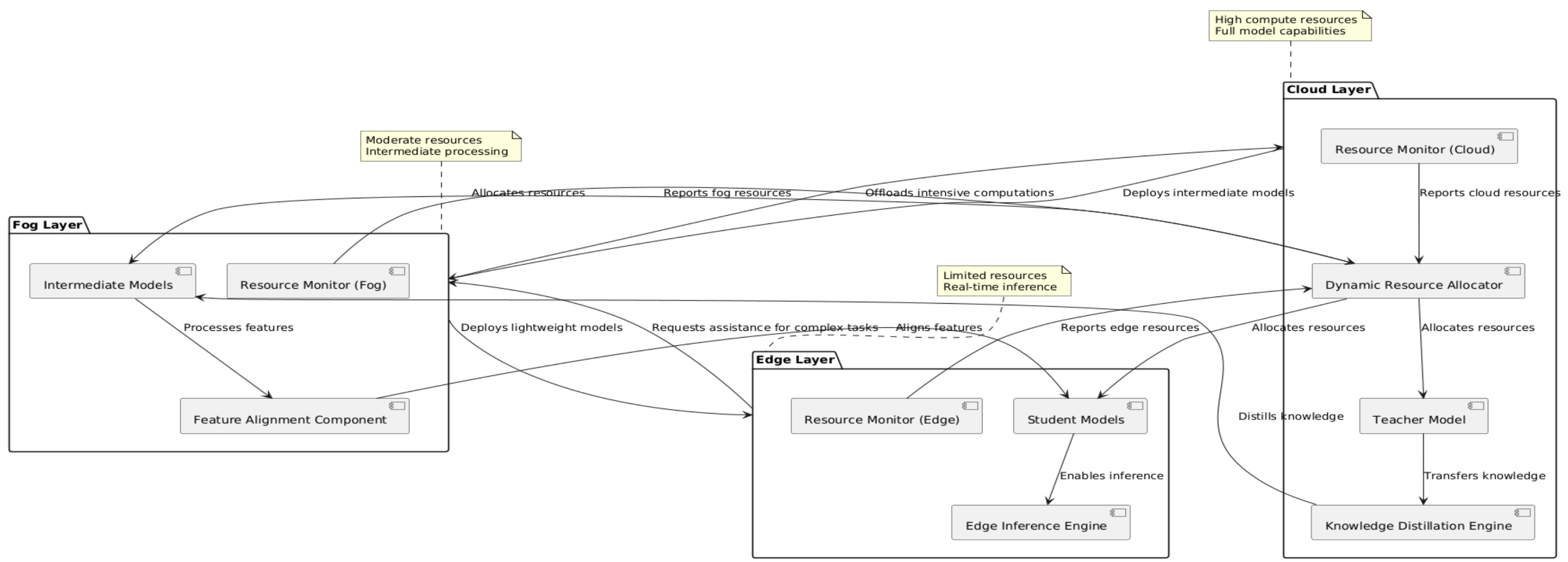
The RA-CMKD framework aims to address these challenges through several key objectives. It develops a dynamic resource allocation mechanism that adapts to hardware limitations, ensuring optimal performance under varying conditions. The framework creates modality-specific compression techniques for visual, audio, and sensor data, preserving essential information while reducing resource requirements. It implements lightweight feature alignment strategies for efficient knowledge transfer between modalities, enabling comprehensive analysis despite resource constraints. The system includes adaptive CMKD loss functions that intelligently balance performance and resource usage based on current conditions. Importantly, the framework enables real-time edge deployment with continuous monitoring capabilities, ensuring sustained performance in dynamic environments. These objectives are validated through comprehensive experimentation across different resource scenarios.

Methodology

The methodology of the RA-CMKD framework centers around three primary innovations. First, the Dynamic Resource Allocation Module continuously monitors CPU, memory, energy, and thermal conditions, implementing a priority-based allocation algorithm that preserves critical modality operations while adaptively distributing resources in response to changing conditions. Second, Modality-Specific Compression Techniques apply structured pruning, quantization, and knowledge-preserving feature extraction for visual data; frequency-domain sparsity and temporal redundancy reduction for audio data; and statistical aggregation and dimension reduction techniques for sensor data. Third, the Lightweight Feature Alignment Strategy employs subspace projection methods for modality-independent feature extraction, adaptive precision control based on feature significance, and optimized contrastive alignment methods for cross-modality information transfer. These components operate within a three-tier hierarchical design spanning cloud, fog, and edge layers, interacting through standardized APIs with dynamic workload distribution based on resource availability.

Model Accuracy Formula: $\mathrm{ACC_m} \times (\frac{\mathrm{LAT_m}}{\mathrm{TAR}})^{w}$

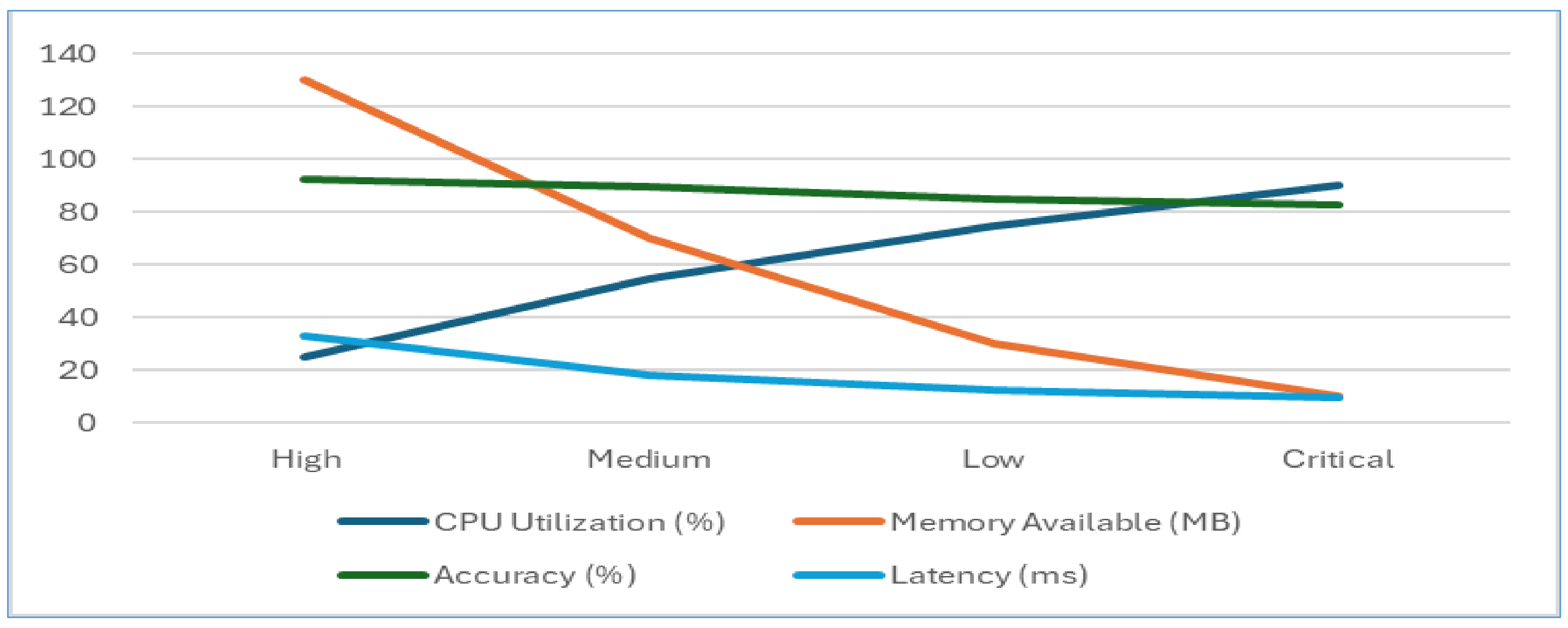
Core Loss Calculation: $L=\alpha(R)\times L_{task}+ \beta(R) \times L_{distill} + \gamma(R) \times L_{align} + \delta(R) \times L_{resource}$



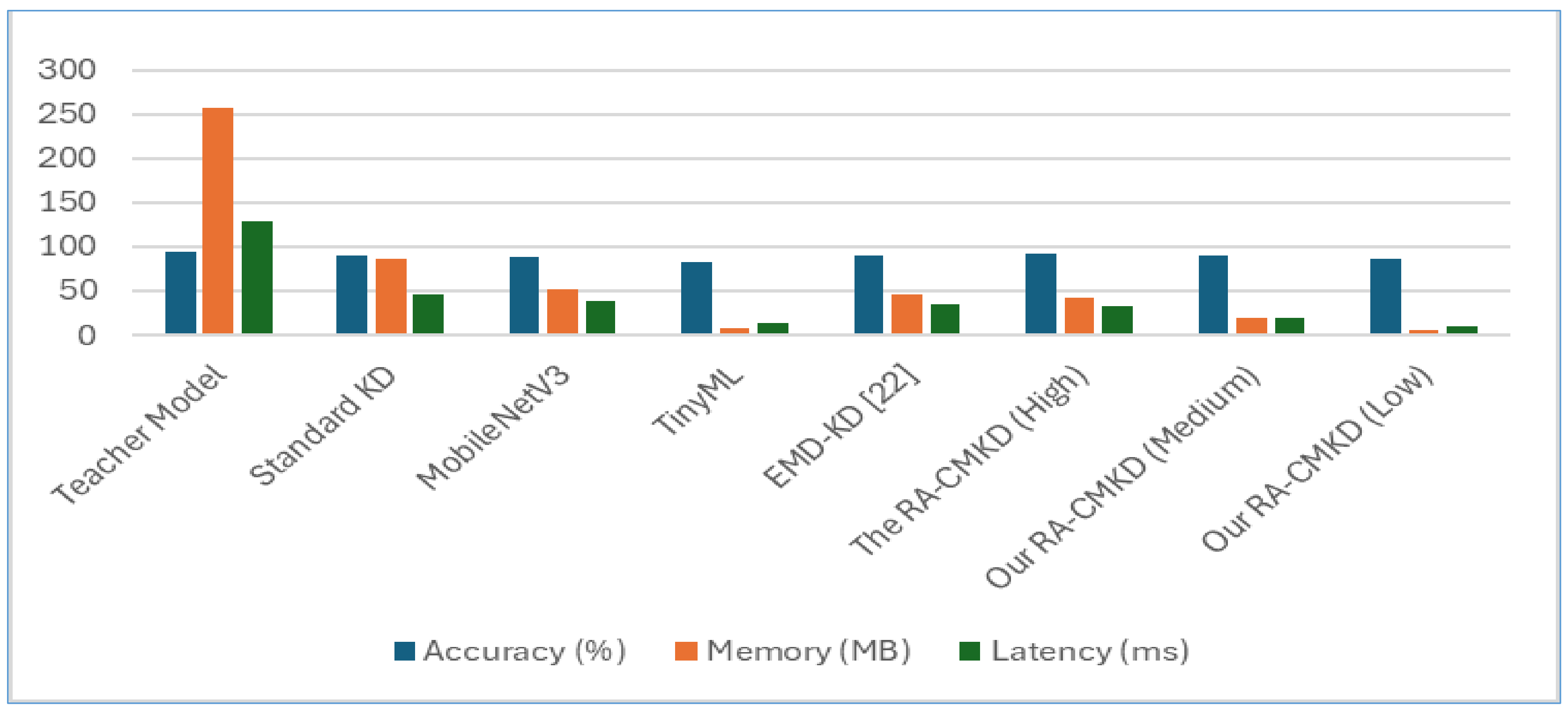
Three-Tier Hierarchical Architecture Fsunctioning

Results

Experimental results with the AV-MNIST dataset in a cloud-fog-edge infrastructure demonstrated the framework's effectiveness across varying resource conditions. The teacher model achieved 94.8% accuracy but required 256.4 MB memory, 128.7 ms latency, and 1250.6 M FLOPs. Under high-resource conditions, RA-CMKD maintained 92.3% accuracy (only a 2.5% reduction) while achieving 83.4% memory reduction, 74.5% latency reduction, and 86.5% computational complexity reduction. Even under critical resource limitations (CPU utilization >90% and available memory <20 MB), the system preserved 82.8% accuracy with just 9.3 ms latency, ensuring real-time performance. Comparative analysis showed that the self-supervised learning approach outperformed EMD-KD and TinyML methodologies. Each component in the three core modules proved essential for system performance, and the framework demonstrated superior accuracy with minimum resource utilization across all experimental setups.



Performance Under Dynamic Resource Constraints



Comparative Study of Proposed Model with existing models and at various Resource Availability Levels

Resource Level	CPU Utilization (%)	Memory Available (MB)	Accuracy (%)	Latency (ms)	Active Modalities
High	20-30	120-150	92.3	32.8	Visual + Audio
Medium	50-60	60-80	89.5	18.2	Visual + Partial Audio
Low	70-80	20-40	85.2	12.5	Visual Only
Critical	90+	<20	82.8	9.3	Visual (Downsampled)

Conclusion

The RA-CMKD framework successfully addresses the challenges of deploying sophisticated multi-modal AI on resource-constrained edge devices. By implementing resource-adaptive mechanisms, modality-specific compression, and lightweight feature alignment, the system maintains high performance while significantly reducing resource requirements. This approach provides an applied edge-device AI deployment framework that advances knowledge distillation science for multi-modal frameworks under resource restrictions. Future work will expand the framework to include additional sensor types, develop techniques for continuous adaptation of deployed models, create more precise resource estimation methods, examine privacy benefits through federated learning integration, and investigate further optimization for specific application domains.

References

[1] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. NIPS Deep Learning and Representation Learning Workshop.

[2] Gupta, S., Hoffman, J., & Malik, J. (2016). Cross Modal Distillation for Supervision Transfer. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2827-2836.