Partitioning Techniques in AI Models for Wireless Network Optimization

Distributed Machine Learning

NAME	BITS ID	CONTRIBUTION
SUBHRANSU MISHRA	2023AC05489	100%
DULAL DAS	2023AC05041	100%
LAKSHMISRINIVAS PERAKAM	2023AC05540	100%
ARCHAN GHOSH	2023AC05042	100%

Part 3: Strategic Analysis and Recommendations: Real-World Recommendations Derived from Experimental Findings

When is Vertical Partitioning More Advantageous?

Vertical partitioning proves highly beneficial in Edge–Cloud distributed AI architectures, particularly when the data modalities or feature sources are heterogeneous:

- Ideal for resource-constrained edge environments, where:
  - Lightweight models operate on real-time features (e.g., Signal Strength, Traffic Volume) near the data source.
  - More complex, high-dimensional user data (e.g., User Count, Device Type) is processed in the cloud or core data centers.
- Applicable when:
  - Input features can be semantically grouped by their origin or latency sensitivity.
  - Communication overhead, data privacy, or regulatory constraints prevent centralized data aggregation.

Our findings show that vertical partitioning, when combined with stacked fusion (meta-modeling), leads to improved predictive performance (lower MAE/RMSE) compared to monolithic, all-in-one models.

When is Horizontal Partitioning Preferable?

Horizontal partitioning is particularly effective in geographically diverse or regionally distributed systems, such as:

- Telecom networks segmented by tower ID, region, or user demographic, where local data exhibits unique characteristics.
- Enables:

  - Localized learning, allowing each model to specialize in the statistical patterns of its region (e.g., Urban vs. Rural).
  - Reduction in centralized training bottlenecks and communication costs, especially beneficial for federated learning scenarios.

In our experiments, the rural model achieved superior MAE and RMSE, while the urban model showed a positive R^2, demonstrating that localized models can outperform global models in capturing regional signal–latency dynamics.

Implications for Telecom Infrastructure and 5G Optimization

- **Optimizing 5G Network Intelligence**

  - **Vertical Partitioning enables layered intelligence:**
    - Edge nodes run fast, low-latency models on infrastructure data.
    - Cloud analytics incorporate historical and behavioural insights for deeper reasoning.

  - **Horizontal Partitioning empowers:**
    - Deployment of region-specific models per tower cluster, capable of adapting to real-time load conditions, user density, and traffic patterns.
    - Enhancement of URLLC (Ultra-Reliable Low Latency Communication) use cases via regional adaptation.

- **Enabling Scalable Edge AI in Wireless Systems**
  - Edge-deployed sub-models reduce reliance on central links and improve autonomy.
  - Distributed model training across towers or edge devices enables scalable, personalized inference pipelines that evolve with regional dynamics.

- **Deployment Challenges**
  - Data Integration: Aligning outputs from independently trained partitions for unified predictions.
  - Communication Latency: Sub-model synchronization may introduce real-time inference delays.
  - Model Consistency: Variability across partitions can cause prediction drift or decision conflict.
  - Resource Constraints: Some edge devices lack the capacity for large models or frequent updates.

- **Strategic Solutions and Design Guidelines**
  - **Model Fusion Mechanisms:** Implement stacked meta-models, weighted averaging, or attention-based ensembling.
  - **Asynchronous Edge-Cloud Inference:** Use real-time edge execution with delayed cloud synchronization.
  - **Federated Learning Pipelines:** Train models across towers or users without moving raw data.
  - **AutoML and Meta-Learning:** Dynamically tailor architectures per region, tower load, or user type.

- **Hybrid Partitioning Strategy:** Combine vertical (feature-wise) and horizontal (sample-wise) approaches.

  Scenario: Combine both strategies for a nationwide 5G deployment.
  - Vertical partitioning used within each region to split real-time and historical processing.
  - Horizontal partitioning used to divide data across towers.
  - Deployment Benefit: Highly distributed, real-time, and resource-efficient modelling setup.

Final Takeaway

Partitioning strategies when aligned with the system's architectural topology and operational constraints can act as a foundational pillar for building scalable, resilient, and real-time AI systems in next-generation wireless infrastructure.

Such architectural modularity empowers telecom providers to:
- Predict and react to latency fluctuations
- Optimize resource allocation
- Personalize service delivery
- And scale seamlessly across diverse geographic regions and edge nodes

This design-first approach is not just an enhancement; it is a necessity for enabling intelligent, decentralized, and adaptive 5G+ networks.