Decentralized Multimodal AI Agents for Resource-Constrained Inferencing or Decisioning

DISSERTATION

Submitted in partial fulfillment of the requirements of the

Degree: M.Tech in Artificial Intelligence & Machine Learning

By

Mathew Kadambatt
2022AA05887

Under the supervision of

Antonio Di Maio
(Postdoctoral Researcher in
Data-Driven Mobile Networked Systems)
&
Prathyusha V
(Senior Technology Engineer)

Under the Mentorship of

Ashwini Chandrashekharaiah
(Senior Data Science Manager)

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
Pilani (Rajasthan) INDIA

(May,2025)

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
SECOND SEMESTER 2024-25

AIMLCZG628T DISSERTATION

Dissertation Title : Decentralized Multimodal AI Agents for Resource-Constrained Inferencing or Decisioning

Name of Supervisor: Antonio Di Maio / Prathyusha Vadde

Name of Student : Mathew Kadambatt

ID No. of Student : 2023AA05887

Courses Relevant for the Project & Corresponding Semester:

1. Distributed Machine Learning (AIMLCZG515)
2. Advanced Deep Learning (AIMLCZG513)
3. MLOps (AIMLCZG523)
4. ML System Optimization (AIMLCZG518)

Abstract

The emergence of intelligent edge ecosystems, comprising distributed, heterogeneous IoT devices, necessitates the design of decentralized AI agents that can collaboratively learn, infer, and make decisions using diverse data modalities under stringent resource constraints. This paper proposes a comprehensive framework for Decentralized Multimodal AI Agents, integrating advances in Federated Learning (FL), Multimodal Machine Learning (MML), and Multi-Agent Deep Reinforcement Learning (MADRL) to enable scalable, privacy-preserving intelligence across resource-constrained environments.

The paper establishes its approach in Multimodal Federated Learning on IoT or similar diverse simulated data, which demonstrates effective cross-device learning with heterogeneous and non-IID multimodal inputs. To enhance adaptability, it incorporates strategies from Multimodal Online Federated Learning with Modality Missing in IoT, allowing agents to perform robust inferencing and decisioning even when certain modalities are absent or dynamically changing. The decentralized coordination and policy optimization aspects are informed by Federated Multi-Agent Deep Reinforcement Learning for Intelligent IoT Wireless Communications, empowering agents to operate autonomously while maintaining collaborative objectives within dynamic wireless networks.

To meet the challenges of communication bottlenecks and edge deployment, it adopts the asynchronous and temporally weighted aggregation strategies from Communication-Efficient Federated Deep Learning, which reduce synchronization overhead and improve convergence in non-ideal network conditions. To further support resource-constrained execution, it incorporates principles from Efficient Deep Learning: A Survey, employing model compression techniques such as pruning, quantization, and knowledge distillation to ensure real-time performance on edge devices.

The work is further guided by A Survey of Multi-Agent Deep Reinforcement Learning with Communication, which offers insights into scalable agent cooperation with minimal inter-agent bandwidth, and Multimodal Machine Learning: A Survey and Taxonomy, which informs the design of modality fusion, alignment, and translation mechanisms.

Collectively, these foundations support a new class of Decentralized Multimodal AI Agents capable of performing adaptive, efficient, and resilient inferencing or decisioning in real-world IoT deployments. It's concluded by highlighting open challenges, including modality-aware aggregation, asynchronous agent collaboration, and efficient decision reasoning under missing data, and outline future research directions for enabling intelligent, decentralized edge-AI systems.

Key Words: Federated Learning; Decentralized AI; Multimodal Machine Learning; Multi-Agent Reinforcement Learning; Edge Intelligence; Resource-Constrained Inference; Asynchronous Model Aggregation; Modality Missing; IoT Systems; Efficient Deep Learning.

Project Work Title

Decentralized Multimodal AI Agents for Resource-Constrained Inferencing or Decisioning

1.1. Purpose:

The purpose of this project is to design and implement a framework for decentralized multimodal AI agents capable of fast and efficient inferencing or decision-making in environments with limited computational, energy, and communication resources. The system leverages advances in federated learning, multi-agent reinforcement learning, and multimodal data processing to enable collaborative, privacy-preserving intelligence across IoT or simulated devices.

1.2. Expected Outcome:

- A scalable architecture for decentralized multimodal AI agents.
- Prototype implementations of agents capable of modality-aware federated learning and inference.
- Performance evaluation under resource-constrained conditions (e.g., CPU-bound, energy-constrained, low-bandwidth).

2. Literature Review:

[1] Multimodal Federated Learning on IoT Data:
Proposes a federated framework for learning from heterogeneous IoT modalities, handling non-IID and multi-sensor data using modality-specific heads and aggregation strategies.

[2] Multimodal Online Federated Learning with Modality Missing in IoT:
Introduces an online FL approach that dynamically adapts to missing or partial modalities, ensuring robust model training despite incomplete input streams.

[3] Federated Multiagent Deep Reinforcement Learning for Intelligent IoT Wireless Communications:
Explores decentralized coordination using multi-agent DRL integrated with FL, emphasizing privacy, scalability, and communication efficiency in wireless IoT networks.

[4] **Communication-Efficient Federated Deep Learning**:
Proposes asynchronous model updates and temporally weighted aggregation to reduce communication overhead and support learning in resource-constrained environments.

[5] Efficient Deep Learning Survey:
Reviews model compression and optimization techniques like pruning and quantization, crucial for deploying AI models on low-resource edge devices.

[6] Survey on Multi-Agent DRL with Communication:
Classifies communication mechanisms in multi-agent RL and their role in enabling cooperative behaviors in distributed intelligent systems.

[7] **Multimodal Machine Learning Survey**:
Provides a taxonomy for handling multimodal data, focusing on fusion, co-learning, and alignment—core challenges for agents handling diverse input streams.

3. Gap Analysis

3.1. Existing Process:

- FL systems typically assume full modality availability and homogeneous data distributions.
- Traditional multimodal ML relies on centralized data processing.
- Multi-agent coordination often assumes ample communication bandwidth and centralized training.

3.2. Limitations:

- Limited support for partial/missing modalities at inference time.
- High communication and synchronization costs for model aggregation.
- Lack of resilience in environments with intermittent connectivity and variable device capabilities.
- Insufficient optimization for real-time inference on resource-constrained edge devices.

4. Justification for Methodology:

A decentralized, federated, and multimodal learning architecture is justified due to:

- **Privacy constraints**: Raw sensor data cannot be centralized due to security/compliance concerns.
- **Heterogeneity**: Devices may have access to different subsets of modalities and varying resources.
- **Scalability**: Centralized AI systems become bottlenecks in large-scale IoT deployments.
- **Adaptability**: The proposed system can operate in dynamic real-world environments where data availability and resource constraints vary in real time.

5. Project Work Methodology:

- **Phase 1: System Design**
  - Define architecture for multimodal federated agents.
  - Incorporate modality-aware model fusion and asynchronous aggregation.

- **Phase 2: Multimodal Development**
  - Develop the base model for each input types
  - Redesign the models for Multimodal basis

- **Phase 3: Model re-design and Agent Development**
  - Integrate compression (quantization/pruning) for efficient model deployment.
  - Improvise the models while preserving the Multimodal logic

- **Phase 4: Deployment and MLOps setup**
  - Setup a GKE cluster on GCP to deploy lightweight models
  - As part of simulation, first set up the models on resource constrained pods

- **Phase 5: Simulation and Benchmarking**
  - Evaluate on IoT/edge simulation platforms or pods with IoT based images
  - Compare performance under various constraints (latency, bandwidth, energy).

- **Phase 6: Improvise the Model Networking & Communications**
  - Validate the model communication and networking as part of Federated learning
  - Analyze and improvise the communication and reduce the latency
  - Implement MADRL to improvise the model performances

- **Phase 7: Real-world Validation (If possible)**
  - Deploy on edge devices (e.g., Raspberry Pi)
  - Assess agent behavior with partial modalities and dynamic task environments

6. Benefits Derivable from the Work:

- **Scalable Edge Intelligence:** Enables real-time decision-making without relying on central servers.
- **Privacy and Security:** Keeps data local to the device, aligning with regulatory standards**.**
- **Fault Tolerance:** Operates even when data or modalities are missing.
- **Efficiency:** Reduces communication and computation overhead via model compression and asynchronous updates.
- **Cross-domain Applicability:** Can be extended to healthcare, smart cities, agriculture, and autonomous systems.

7. Additional Details:

- **Tools/Frameworks**: PyTorch, TensorFlow Federated, Kubernetes, Docker
- **Future Extensions**:
  - Hierarchical FL with agent clustering.
  - Multilingual or multi-task agents.
  - Integration with blockchain for federated trust mechanisms.

1. Broad Area of Work:

The broad area of this research lies at the intersection of:

- **Artificial Intelligence (AI) and Machine Learning (ML) in MultiModal**: Specifically in **Multimodal Learning**, **Federated Learning (FL)**, and **Multi-Agent Systems**.
- **Edge and IoT Computing**: Focusing on AI deployment in **resource-constrained, distributed, and privacy-sensitive environments**
- **Efficient AI Systems**: Emphasizing **communication-efficient**, **computation-light**, and **adaptive AI agents** that operate under real-world constraints.

It contributes to current research trends in **Decentralized AI**, **MultiModal ML**, **On-device Learning**, **Cross-device Collaboration**, and **Edge-centric Intelligence**

2. Objectives

The objectives of my project are as follows:

- **Design a Federated Learning Framework for Multimodal Data**
  Develop a decentralized learning system where agents collaboratively train models using **heterogeneous and distributed multimodal data** (e.g., visual, audio, sensor, or textual inputs), without sharing raw data.
- **Enable Intelligent Decision-Making in Resource-Constrained Environments**
  Create AI agents that can **infer, reason, and act** efficiently under constraints such as limited computation, memory, energy, or intermittent connectivity.
- **Handle Missing or Incomplete Modalities in Real-Time**
  Build models that can dynamically **adjust to the presence or absence of certain modalities** (e.g., missing data, sensor unavailable) during both training and inference.
- **Develop Communication-Efficient Learning and Aggregation Techniques**
  Incorporate methods like asynchronous model updates**,** temporally weighted aggregation, and compression-aware communication to minimize bandwidth and latency overhead.
- **Optimize AI Models for On-Device Inference**
  Apply techniques such as model pruning, quantization, and knowledge distillation to make models lightweight and suitable for edge devices.
- **Validate the Framework on Realistic IoT and Edge Platforms**
  Test and evaluate the developed system on physical or emulated edge devices like Raspberry Pi, Jetson Nano, or IoT gateways under virtual operating conditions, if difficult to simulate on real-world operating conditions.
- **Integrate Multi-Agent Deep Reinforcement Learning (MADRL)**
  Facilitate cooperative behavior among distributed agents using **reinforcement learning strategies**, allowing agents to learn optimal policies with minimal direct communication as the last extension after major above major objective
- **Contribute Research Insights and Practical Guidelines**
  Provide theoretical and empirical insights into building decentralized, privacy-preserving, multimodal AI agents for the edge, and contribute to open-source benchmarks or toolkits

3. Scope of Work

The project aims to:

1. **Design a Decentralized Learning Architecture**
    - Architect federated, agent-based systems that operate collaboratively without centralized data collection.
    - Enable agents to process **multimodal input data** (e.g., audio, visual, sensor signals) on-device.
2. **Incorporate Robust Multimodal Learning Mechanisms**
    - Allow agents to learn from and make decisions with **incomplete, missing, or corrupted modalities**.
    - Implement modality-aware aggregation and late fusion strategies within the federated pipeline.
3. **Develop Multi-Agent Coordination Capabilities**
    - Leverage **deep reinforcement learning** and **communication-aware collaboration** strategies to train agents to learn optimal behaviors.
4. **Optimize for Resource Constraints**
    - Implement model compression techniques (e.g., pruning, quantization, distillation) and asynchronous communication schemes to support:
        - **Low power**,
        - **Limited bandwidth**, and
        - **Intermittent connectivity** environments.
5. **Simulate and Benchmark the Framework**
    - Evaluate the performance of the proposed system using synthetic and real-world datasets.
    - Measure accuracy, latency, communication cost, energy consumption, and robustness to missing modalities.
6. **Deploy and Validate on IoT/Edge Devices**
    - Demonstrate real-time decisioning capabilities on actual devices like Raspberry Pi or NVIDIA Jetson Nano under simulated edge conditions.
7. **Produce Research Outputs and Future Guidelines**
    - Publish findings on how **decentralized multimodal AI agents** can be extended to various application domains.
    - Outline challenges and roadmap for federated learning with multimodal, asynchronous, and multi-agent settings.

4. Detailed Plan of Work

Sno.	Tasks/Subtasks	Start Date - End Date	Planned Duration (weeks)	Specific Deliverable
1	Research, Analysis and Scope Definition	Week 1 - Week 2 (May 16th)	2	Research, Objectives/Scope of work analysis and Abstract submission
2	Phase 1: System Designing	Week 3 - Week 4 (May 30th)	2	Designing the model to incorporate FL and Multimodal logic
3	Phase 2 & Phase 3: Multimodal development and Redesign	Week 5 - Week 6 (June 13th)	2	Develop the model and later compress it based on Literature Ref [5]
4	Phase 4: Infra setup and Deployment	Week 7 - Week 8 (June 27th)	2	GKE(k8s) infra setup, MLOps configuring & deployment
5	Phase 5:Simulation using raw data	Week 9 - Week 10 (July 11th)	2	Simulate the scenario. Mid Sem submission and Evaluation
6	Phase 6: Improvise the model networking and communication	Week 11 - Week 12 (July 25th)	2	Decentralized federated learning improvisations from network point of view
7	Phase 7: (Optional) Implement real time scenario	Week 13 - Week 14 (Aug 8th)	2	Simulate IoT device image & communication in G and analyze or use real world IoT device.
8	Optimization, Pending tasks & Final Evaluation	Week 15 - Week 16 (Aug 22nd)	2	Implement Optimizations & Complete pending tasks & Final Report submission & Viva

5. Literature References:

The following are referred journals from the preliminary literature review.

[1] Chen, J., Sun, Q., & Jin, Y. (2019**). Communication-Efficient Federated Deep Learning with Asynchronous Model Update and Temporally Weighted Aggregation.** *arXiv preprint arXiv:1903.07424*. https://doi.org/10.48550/arXiv.1903.07424

[2] Wang, H., Liu, X., Zhong, X., Chen, L., Liu, F., & Zhang, W. (2025). **Multimodal Online Federated Learning with Modality Missing in Internet of Things.** *arXiv preprint arXiv:2505.16138*. https://doi.org/10.48550/arXiv.2505.16138

[3] Chen, X., Hu, Y., He, J., & Xu, C. (2021). **Multimodal Federated Learning on IoT Data**. *arXiv preprint arXiv:2109.04833*. https://doi.org/10.48550/arXiv.2109.04833

[4] Li, X., Liang, Y., Xu, Y., & Guo, W. (2024). **Federated Multiagent Deep Reinforcement Learning for Intelligent IoT Wireless Communications: Overview and Challenges**. *IEEE Communications Magazine*. https://doi.org/10.1109/MVT.2024.3451191

[5] Menghani, G. (2021). Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better. *arXiv preprint arXiv:2106.08962*. https://doi.org/10.48550/arXiv.2106.08962arXiv+1arXiv+1

[6] Zhu, C., Dastani, M., & Wang, S. (2022). **A Survey of Multi-Agent Deep Reinforcement Learning with Communication**. *arXiv preprint arXiv:2203.08975*. https://doi.org/10.48550/arXiv.2203.08975

[7] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2017). **Multimodal Machine Learning: A Survey and Taxonomy**. *arXiv preprint arXiv:1705.09406*. https://doi.org/10.48550/arXiv.1705.09406