Midterm Report: Adaptive Multimodal Deep Network for Real World Data

Cheng-Hsiu (Alan) Hsieh, Ting-Yu Yeh, Chin-Yi (Daniel) Lee

Motivation and Objectives

- Objectives
  - Design a multi-modal sensing embedded device
  - Dynamically allocate compute resources based on each modality's quality (QoI) to ensure real-time performance.
- Where to apply?
  - Outdoor smart garage
  - Remote vehicle control
- Goals
  - Multi-modality sensing
  - Limited resources
  - Minimum power

Technical Approach and Novelty

- Current Approach
  - ADMN (Adaptive Depth Multimodal Network)
  - Dynamically allocate resources (for different modalities)
  - Tested on synthesized noise

- Our Approach
  - Implement on edge device (scale down model size)
  - Validate its feasibility with real world data and noise

ADMN: Wu, J., Yang, K., Kaplan, L., & Srivastava, M. (2025). ADMN: A Layer-Wise Adaptive Multimodal Network for Dynamic Input Noise and Compute Resources. ***arXiv preprint arXiv:2502.07862*****. doi: 10.48550/arXiv.2502.07862**

Methods (1) - Overview

1. Data Collection & Preprocessing
   (Collect Clean & Corrupted Data)

2. Model Fine-tuning & Optimization
   (2 Stage Training: Baseline & Adaptive)

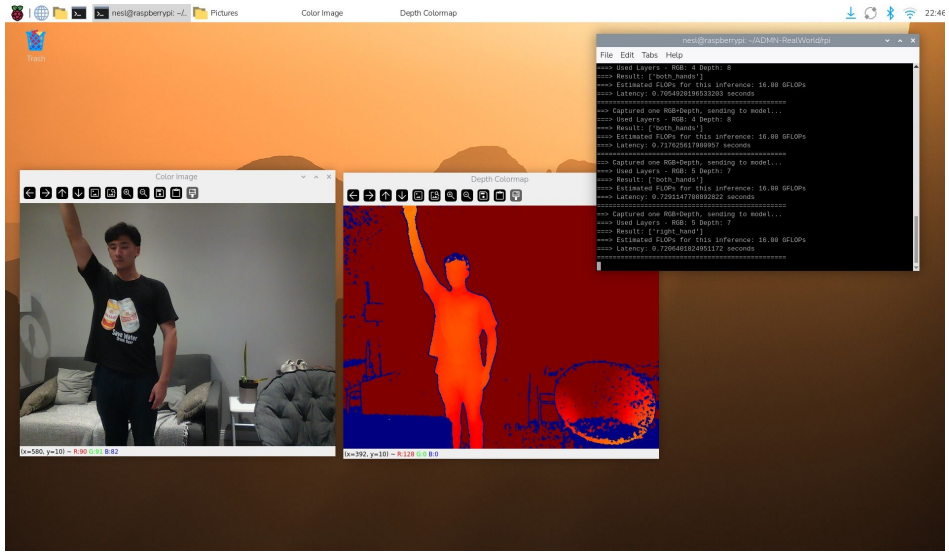
3. Edge Deployment & Real-Time Inference
   (Load Model on Rpi 5 for real-time inference)

3 Data Conditions:
- clean
- depth_occluded
- low_light

4 Gesture Classes:
- standing
- left_hand
- right_hand
- both_hands

Total Samples:
240 (80 per condition)

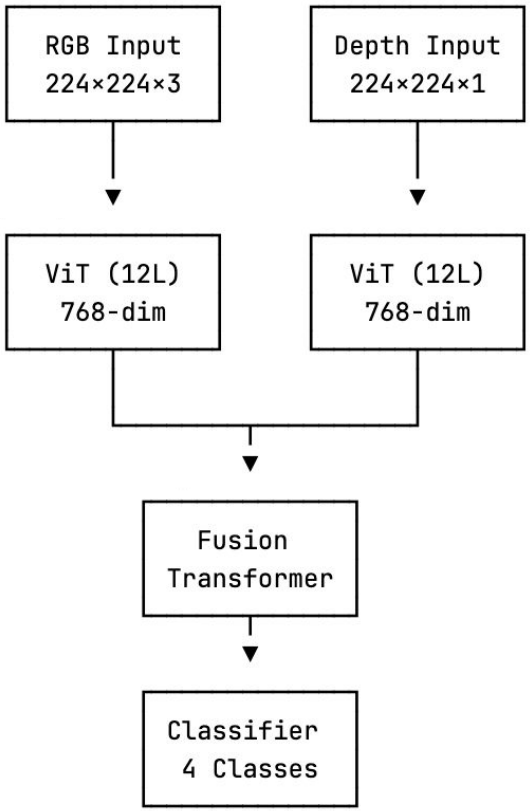
Resolution:
- RGB (224×224×3)
- Depth (224×224×1)



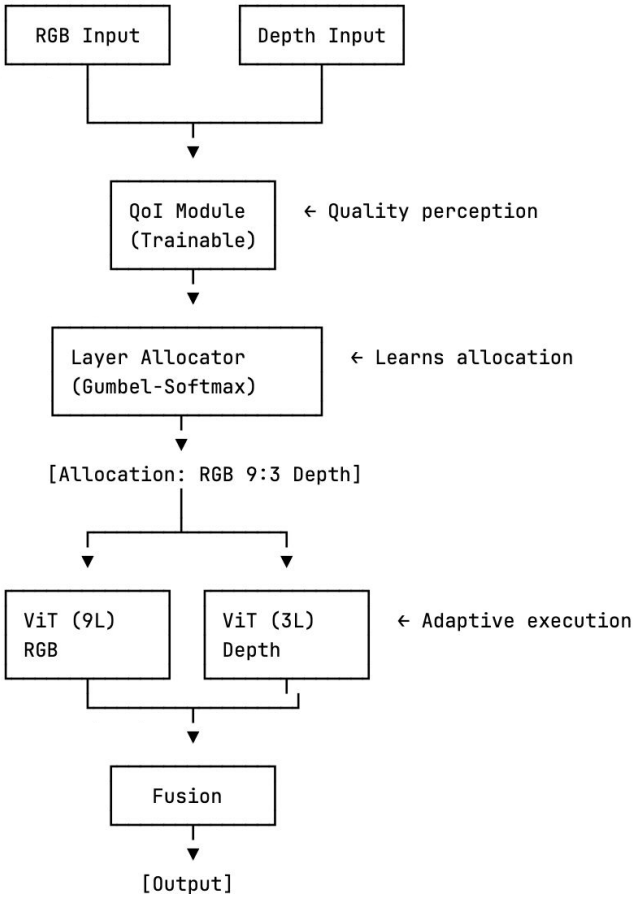
Real-time adaptive inference demo

Methods (2) - Two-Stage Training

Stage 1: LayerDrop Finetuning



Stage 2: Controller Training



Evaluation and Metrics (1)

Performance Summary

Model	Accuracy	Adaptation Strategy
Stage 1 (Baseline)	95.83%	Fixed: 12 layers per modality
Stage 2 (Adaptive)	95.83%	**Dynamic allocation based on quality**

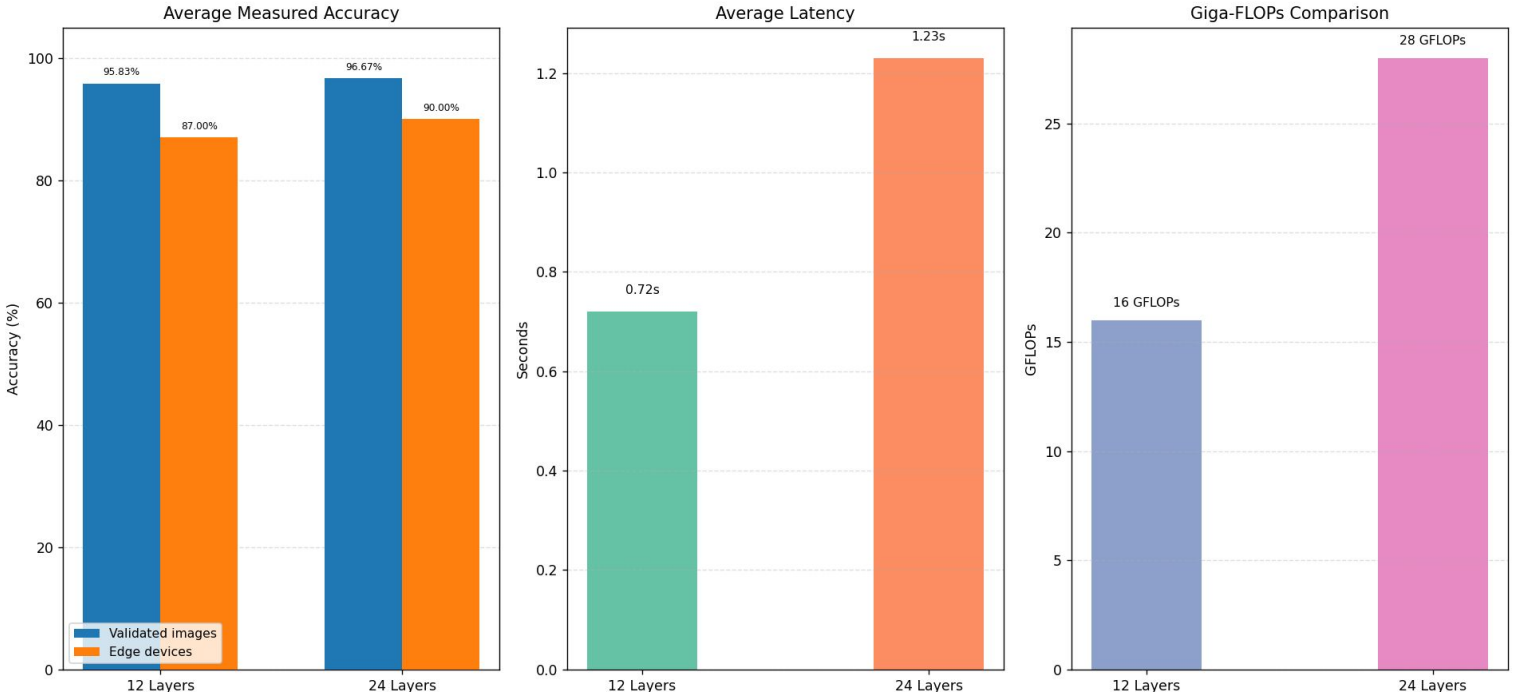
Adaptive Allocation Strategy

Our Stage 2 model successfully learned corruption-aware allocation:

Corruption Type	RGB Layers	Depth Layers	Strategy
Clean	5.3 / 12	6.7 / 12	Balanced allocation ⚖️
Depth Occluded	9.1 / 12	2.9 / 12	Allocate to RGB 🔴
Low Light	2.2 / 12	9.8 / 12	Allocate to Depth 🔵

Evaluation and Metrics (2)

- **Baseline (24 layers):** Fixed allocation of 12 RGB layers and 12 Depth layers
- **Our Model (12 layers):** Dynamically allocates 12 layers between RGB and Depth based on QoI (Quality-of-Information)
- **Evaluation setup:** Tested on different conditions



Current Status and Next Steps

- Completed
  - Implement ADMN architecture for RGB-D gesture recognition
  - Achieve adaptive layer allocation based on input quality
  - Deploy on edge device (RPi 5) for real-time inference

- Next Steps
  - Real-time inference with GPIO actions
  - Sleep mode with event-based wake-up
  - Real-time monitoring webpage
  - Benchmark generation & comparison