AAI-521 – Team Project Status Update Form

Team Number: Gangadhar Singh Shiva, Akshobhya Rao BV, Nagarajan Mahalingam

Team Leader/Representative: Gangadhar Singh Shiva,

Full Names of Team Members:

1. Gangadhar Singh Shiva
2. Nagarajan Mahalingam
3. Akshobhya Rao BV

Project Title

A Scalable Agentic AI Framework for Multi-Modal Video Classification,Summarization and Analysis

Short Description of Your Project and Objectives:

The system utilizes a multi-modal, six-step LangGraph pipeline, integrating vision, audio, and real-time contextual grounding. Core components include a ResNet18 backbone fine-tuned using LoRA (Low-Rank Adaptation) for parameter-efficient classification, and a YOLOv8n model combined with the Structural Similarity Index (SSIM) for efficient visual summarization. The architecture features an autonomous workflow that dynamically processes raw video through keyframe extraction, optional audio transcription, and real-time news retrieval. This demonstrates a robust fusion of computer vision and natural language processing techniques within a stateful, agent-based environment. Initial validation confirms the successful functional integration of all primary models and highlights key areas for immediate parameter optimization, specifically regarding the high keyframe extraction rate.

Introduction

The rapid expansion of multimedia data—especially videos—has led to the need for intelligent systems capable of not only identifying actions but also explaining and contextualizing them. Traditional deep neural networks (DNNs) like CNNs perform well at classification but operate as "black boxes" lacking interpretability and autonomy.

This project introduces Agentic AI, an approach where the model learns from visual data, classifies unseen inputs, evaluates its confidence, decides whether retraining or summarization is required, generates natural language descriptions, and links outputs to real-world news and sentiment.

The **Kaggle Human Activity Recognition dataset** provides rich, labeled time-series and video data representing six distinct human activities:

- Walking
- Walking Upstairs
- Walking Downstairs
- Sitting
- Standing
- Laying

By combining **ResNet-18 + LoRA** for efficient fine-tuning with **LangGraph** for agentic decision flow, **YOLO8n** for video extraction and highlighting, the system demonstrates multi-modal intelligence and contextual understanding beyond simple classification

By combining ResNet-18 + LoRA, Yolo*n with LangGraph, the system achieves efficiency, transparency, and autonomy, marking a significant advancement in intelligent video understanding.

Goals

To develop a **self-adaptive multimodal agentic AI system** capable of recognizing, summarizing, and contextualizing human activities using a combination of vision and language models.

2.2 Specific Objectives

The objective is to create an autonomous system that not only accurately classifies video content but also efficiently generates a concise summary and grounds the content within real-time external context.

Are you using and practicing GitHub as a code hosting platform for version control and collaboration? If yes, provide the link here: **https://github.com/gshiva1975/AAI-521-Project/**

How many times have your members met in the last two weeks? 3 Times

Project Objectives

- Integrate the **Kaggle HAR dataset** for human activity classification.
- Design a **custom VideoDataset class** for frame-based feature extraction using **YOLOv8n.**
- Fine-tune **ResNet-18** with **LoRA** to improve efficiency and generalization.
- Build an **agentic LangGraph workflow** that handles classification, retraining, summarization, and sentiment analysis.
- Retrieve **contextual news** related to detected activities using **Google News RSS**.
- Perform **sentiment analysis** on the retrieved headlines to assess social or emotional tone.

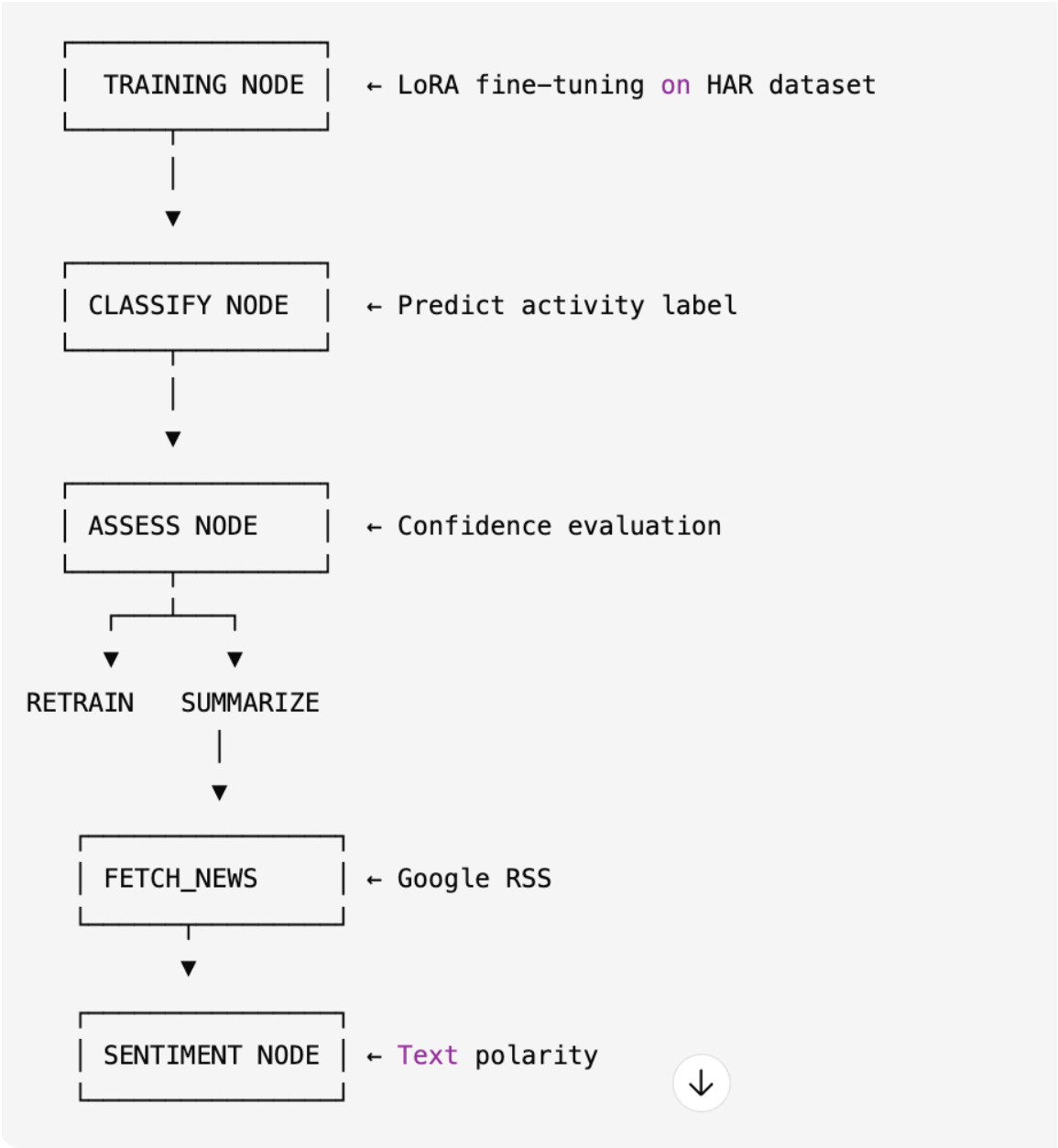
System Design

- The system is composed of three main layers: (1) the Perception Layer for video data handling and feature extraction, (2) the Reasoning Layer for LangGraph-based decision flow, and (3) the Context Layer for text summarization, news retrieval, and sentiment

analysis. The architecture integrates multiple AI disciplines into a unified agentic framework.

- **Architectural Flow**

  The flow begins with model training using LoRA fine-tuning, followed by video classification. The confidence of predictions is assessed; if low, retraining is triggered, else summarization is performed with YOLO8n. The summary triggers contextual news fetching and sentiment analysis, creating an autonomous cycle of perception and reasoning.


TRAINING NODE
← LoRA fine-tuning on HAR dataset
CLASSIFY NODE
← Predict activity label
ASSESS NODE
← Confidence evaluation
RETRAIN
SUMMARIZE
FETCH_NEWS
← Google RSS
SENTIMENT NODE
← Text polarity

Work Breakdown Structure (WBS)

List the specific contributions that each team member is providing for the Final Team Project in the table below.

- **NOTE:** ALL students on the team should contribute equally to the Final Team Project.

Gangadhar Singh Shiva	Nagarajan Mahalingam	Akshobhya Rao BV
Create Github Repo Data Collection & Cleaning	Data Collection & Cleaning	Data Collection & Cleaning
Exploratory Data Analysis	Exploratory Data Analysis	Exploratory Data Analysis
Langgraph Agentic AI Implementation	Yolo8n Summarization Implementation	Resnet-18, Lora Model Training and Classification
Integration of Code	GoogleNEWS - Text & Context summarization	Integration of Code
Model Training & Testing, Documentation	Model Training & Testing, Documentation	Model Training & Testing, Documentation

Comments/ Roadblocks:

This project balances AI innovation (LLM-driven reflection, memory) with practical video analysis.