Scalable Agentic AI Agent for Video Analysis

with LoRA Fine-Tuning and LangGraph Orchestration

Gangadhar S Shiva, Akshobhya Rao BV , Nagarajan Mahalingam

November 2, 2025

Abstract

This report details the implementation of a scalable, agentic Artificial Intelligence (AI) system designed for comprehensive video analysis, integrating classification and natural language processing (NLP) summarization. The core methodology employs **Low-Rank Adaptation (LoRA)** for memory-efficient fine-tuning of a pre-trained **ResNet18** model for video action classification. The entire workflow is orchestrated by a **LangGraph**-based agentic pipeline, which includes sequential tasks like video classification, audio extraction, transcription, key frame summarization, and external research/news fetching based on the classified content. This approach demonstrates a powerful and resource-efficient framework for processing large video datasets.

1 Introduction

The increasing volume of video data necessitates efficient and automated processing solutions. Traditional full fine-tuning of deep learning models for video classification is computationally expensive and memory-intensive. This project addresses these challenges by combining three key technologies: a deep learning model (ResNet18) for video classification, LoRA for parameter-efficient fine-tuning, and the LangGraph framework for orchestrating a multi-step, agentic analysis pipeline. The final system is a modular agent capable of classifying video content, generating a concise video summary, and conducting related research.

2 Methodology

2.1 Video Classification with LoRA Fine-Tuning

Base Model and Dataset The video classification task utilizes a pre-trained **ResNet18** model, adapted for video inputs by processing stacked frames. The model was fine-tuned on a video dataset structured for action recognition, containing classes such as `CricketShot`, `PlayingCello`, `Punch`, and `ShavingBeard`.

Low-Rank Adaptation (LoRA) To ensure scalability and reduce resource requirements, **LoRA** was applied to the ResNet18 model using the PEFT library. LoRA significantly decreases the number of trainable parameters by injecting small, low-rank matrices ($B \times A$) into the pre-trained weights (W_0) of the attention layers. This method allows for efficient fine-tuning without updating the massive base model weights, W_0.

2.2 Agentic Pipeline using LangGraph

A state-based graph framework, **LangGraph**, was used to build a robust and fault-tolerant agent capable of sequential video analysis tasks. The agent's state (`SummaryState`) maintains crucial information, including the video path, predicted label, and intermediate results like the transcript and news articles.

The pipeline consists of the following orchestrated nodes:

1. **Audio Extraction:** Extracts the audio track from the input video (`.avi`) using `moviepy` and saves it as a WAV file.

2. **Transcription:** Converts the extracted audio to text using the `speech_recognition` library (Google Web Speech API).
3. **Summary Video Creation:** Identifies key segments (by extracting frames at regular 15-frame intervals) and concatenates 1-second clips from those segments to produce a concise summary video.
4. **Label Processing (Lemmatization):** Processes the model's predicted label (e.g., `CricketShot`) into a searchable keyword (e.g., `cricket`).
5. **News/Research Fetching:** Conducts an external search using the processed keyword to fetch related news articles, providing real-time contextual information.

3 Results and Discussion

3.1 Video Classification Performance

The LoRA fine-tuned model was evaluated on a test video, `v_CricketShot_g01_c03.avi`, and successfully classified the action:

Predicted Label: CricketShot

This result validates the effectiveness of the LoRA fine-tuning approach in adapting the pre-trained ResNet for domain-specific action classification with minimal parameter updates.

3.2 Agentic Pipeline Execution Trace

The LangGraph agent executed the full analysis pipeline for the sample video. The trace revealed the following key outcomes:

- **Audio and Transcription:** Audio extraction was successful, but the transcription step failed, returning `sr.UnknownValueError` (and thus, "No speech detected in video"). This indicates the video was primarily visual, lacking discernible speech content.
- **Summary Video:** A summary video was successfully generated by concatenating 1-second clips from 6 identified key frames. The final summary video duration was 6 seconds, demonstrating the successful extraction of representative visual segments.
- **Contextual Analysis:** The predicted label `CricketShot` was successfully processed to the search query `cricket`. The agent then successfully fetched 100 related news articles, proving the agent's capacity for integrated video analysis and external research.

4 Conclusion

This project successfully implemented a scalable and modular AI system for video analysis, named the "Scalable Agentic AI Agent". Key achievements include:

- Demonstrating the efficacy of **LoRA Fine-Tuning** for efficient and memory-saving adaptation of a video classification model.
- Orchestrating a complex, multi-modal video analysis pipeline using **LangGraph**, integrating visual (classification, summarization) and audio (transcription) processing with external research.
- The agent successfully generated a key-frame summary and provided relevant contextual news research, proving the end-to-end functionality of the agentic framework.

Future work could focus on integrating more robust visual analysis methods for content summarization when speech is absent.

References

1 GangadharSShiva. Copy_of_working_pretrainemodelGangadharSSingh_Assingment_project.ipynb. (Uploaded Jupyter Notebook).

2 GangadharSShiva. project_AAI_521_1.ipynb. (Uploaded Jupyter Notebook).