Course Name: Computer Vision

Weekly Report: 7

Group Name: XYZ

Submitted to faculty:

Mehul Raval

Date of Submission:

12 Mar 2025

Student Details

Roll No.	Name of the student	Name of the program
AU2240106	Meet Rathi	B.Tech in CSE
AU2240160	Harsh Panchal	B.Tech in CSE
AU2240153	Aditya Agarwal	B.Tech in CSE

Table of Contents.

WORK DONE THIS WEEK

The Jetson Orin AGX platform received modifications to its codebase and work for deploying TensorRT models to execute them.
The main work of this week included developing Jetson-specific functionality codes and deploying TensorRT models for execution.

Key Activities:

- The LoRA-tuned CLIP model succeeded in its TensorRT conversion to reach both quicker execution speed and reduced memory needs during Jetson Orin AGX inference tasks.
- The Jetson Codebase Refactoring handled the critical codebase to make it integrate with Jetson SDKs which included the following modifications:
    - The inference calls that demanded high processing speed required a replacement which led to the adoption of TensorRT runtime execution.
    - The logical procedures that process data need to be altered to work effectively with the Jetson device specifications.
    - Edgeworth outcomes were introduced when original dependencies met their replacement with updated alternatives.
- CUDA & TensorRT Integration: Integrated CUDA streams and TensorRT engine execution into the inference pipeline for low-latency performance.
- Our team adapted the inference operation performance by using specific precision settings (FP16) along with adjusted memory

management and created suitable batch size parameters aimed at optimizing Jetson Orin AGX's GPU memory usage.

- Jetson underwent basic person retrieval testing by applying test inputs to achieve operational efficiency during edge operations.

WORK TO BE DONE NEXT WEEK

System deployment on Jetson Orin AGX begins after we focus on final optimization tests and prepares the system for stable operation.

Key Objectives:

Pipeline Optimization

- Besides I/O overhead reduction and CUDA execution optimization the system achieves higher performance through precision optimization which includes FP16/INT8 for lower latency and memory usage.

Dataset Evaluation

- The assessment of model performance and real-world accuracy for benchmark datasets (RSTP-ReID, CUHK-PEDES, ICFG-PEDES) becomes necessary for the deployed model testing.

Performance & Power Profiling

- Testing of runtime performance measurement data including GPU management and power usage and execution speed must be performed to identify edge computing effectiveness.

Final Deployment Setup

- The project needs to create clean modular deployment scripts which target Jetson systems to ensure efficient dependable system execution.

Documentation & Reporting

- The project documentation should conclude with results of performance testing as well as setup guides and deployment procedure.