Ahmedabad
University

Course Name: Computer Vision

Weekly Report: 6

Group Name: XYZ

Submitted to faculty:

Mehul Raval

Date of Submission:

5 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

Model Optimization with Pretrained Architectures

Our focus this week was on CLIP-based person retrieval model optimization through evaluation of pretrained backbones to determine their effects on accuracy-speed-size balance before our planned Jetson deployment.

Key Activities:

- We conducted experimental evaluations that used various pretrained variants including ViT-B/32 and RN50x4 and others.
- LoRA techniques applied Parameter-Efficient adaptations through the use of Low-Rank Adaptation (LoRA) to decrease trainable parameter numbers while maintaining retrieval performance.
- The research team observed model inference times and memory usages of various backbones while searching for an efficient performance setup through their inference profiling activities.
- The potential of Jetson hardware constraints was evaluated through initial tests which combined model pruning techniques with dynamic quantization procedures.
- A verification step confirmed that every optimized model functions properly within the Jetson Orin AGX hardware constraints especially regarding its memory capacity and TensorRT conversion properties.

WORK TO BE DONE NEXT WEEK

The TensorRT conversion process enables the deployment of the model on Jetson Orin AGX devices.

- The selected model needs TensorRT optimization to achieve more efficient performance on edge hardware for inference use.

Refactor codebase for Jetson compatibility

- The project requires a modification of existing code to incorporate Jetson-specific libraries TensorRT together with CUDA and streamline GPU operations.

Different inference tests will run while measuring system performance metrics

- Post-deployment testing of the model should focus on speed as well as power efficiency along with retrieval accuracy within real-time requirements.