

Smart Edge lens and Cloud Optimization

1st Tejumal Movin

Department of Computer Science
Rishihood University, Sonipat, India
sansar.t23csai@nst.rishihood.edu.in

2nd Sahil Sarawgi

Department of Computer Science
Rishihood University, Sonipat, India
sahil.s23csai@nst.rishihood.edu.in

3rd Abhishek Meena

Department of Computer Science
Rishihood University, Sonipat, India
abhishek.m23csai@nst.rishihood.edu.in

Abstract—This work focuses on building a Smart Edge Lens system capable of performing real-time anomaly detection using Vision Transformers for traffic monitoring. Instead of transmitting raw high-volume video streams, frames are analyzed on the edge for efficiency. To further optimize performance, model quantization techniques are considered for reducing inference latency. Initial implementation includes dataset preparation, frame extraction, base Vision Transformer experimentation and performance observation in Google Colab.

Index Terms—Vision Transformer, Edge AI, Quantization, Traffic Dataset, Surveillance, Deep Learning

I. INTRODUCTION

Traffic monitoring plays a crucial role in smart city development. Traditional surveillance systems upload continuous CCTV video streams to centralized servers, which leads to bandwidth overload, increased latency, and high infrastructure cost. To overcome this limitation, edge-based inference is required where analysis occurs near the data source.

In this project, we aim to develop a **Smart Edge Lens** system that processes traffic frames on the edge using Vision Transformers (ViT/DeiT/MobileViT). By performing anomaly detection locally and compressing event information, the system becomes faster and more scalable. The work completed so far includes dataset selection, preprocessing, baseline model setup in PyTorch, and exploration of quantization for deployment feasibility.

II. LITERATURE REVIEW

Smart surveillance systems traditionally depend on cloud computation, but increasing video density makes raw data transfer inefficient. Digital Twin frameworks aim to represent real-world systems virtually, but most implementations rely heavily on cloud backends. Our work focuses on the local/edge processing stage, improving efficiency before cloud involvement.

A. Vision Transformers for Image Understanding

Dosovitskiy et al. [1] introduced the Vision Transformer (ViT), treating an image as a sequence of patches fed through a transformer network. Unlike CNNs, ViT enables global feature attention, which is beneficial in traffic scenarios where events occur at multiple spatial regions. In our project, video frames are resized, patched and passed into a ViT-based model for feature extraction and classification.

Touvron et al. [2] developed DeiT, improving training efficiency using knowledge distillation, making ViTs feasible

even without massive datasets. This supports our goal of using publicly available highway datasets for anomaly recognition. DeiT-Tiny serves as a lightweight architecture suitable for Colab-based training.

Howard et al. [3] introduced MobileViT, combining convolution and transformer layers for mobile-friendly operation. This is crucial for edge hardware like Raspberry Pi or Jetson Nano where resources are limited. Based on these works, we began implementing ViT-based processing pipelines using PyTorch.

B. Quantization for Edge Deployment

To deploy deep models on edge devices, reducing compute cost is mandatory. Quantization converts floating-point weights to lower resolution (INT8), reducing memory footprint up to 4x while maintaining accuracy. Research shows INT8 inference retains over 95% of original performance, making it a suitable optimization method.

In our current work, we have prepared the model export pipeline for ONNX conversion, enabling further TensorRT-based quantization. The aim is to decrease inference time and achieve real-time processing on resource-limited platforms.

C. Dataset Processing and Current Implementation

The dataset used consists of highway traffic videos, suitable for detecting congestion or abnormal vehicle activity. Work completed so far includes:

- Video frame extraction using OpenCV.
- Conversion of frames into input tensors for ViT models.
- Implementation and experimentation of baseline ViT in Google Colab.
- Accuracy observation and preparation for quantization-based deployment.

Preprocessing ensures fixed frame resolution, normalization and batch feeding for model training. This enables efficient backbone selection and provides foundation for anomaly recognition pipeline.

III. CONCLUSION

This paper reviewed baseline work on Vision Transformers, lightweight model design and quantization for edge inference. Our implementation currently includes video dataset handling, frame processing and ViT experimentation inside Google Colab. Future goals include INT8 quantization, performance

benchmarking, and integration of event-based filtering for real-time monitoring. As development progresses, the model will be optimized for deployment on edge devices for scalable city surveillance.

REFERENCES

- [1] Dosovitskiy, A. et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2020.
- [2] Touvron, H. et al., "Training Data-Efficient Image Transformers & Distillation Through Attention," ICML, 2021.
- [3] Howard, A. et al., "MobileViT: Lightweight Vision Transformer for Mobile Devices," ICLR, 2021.
- [4] Tao, F., Zhang, H., Liu, A., & Nee, A., "Digital Twin in Industry: State of the Art," IEEE Transactions on Industrial Informatics, 2019.
- [5] NannyML & River., "Open-Source Frameworks for Concept Drift and ML Monitoring," GitHub, 2023.