CRITICAL SCENARIO DETECTION FROM REAL WORLD CAMERA

Khoa Vo
University of Wuppertal
Wuppertal, Germany 42117
ngoc.vo@uni-wuppertal.de

Adwait Chandorkar
University of Wuppertal
Wuppertal, Germany 42117
chandorkar@uni-wuppertal.de

January 12, 2025

ABSTRACT

In this project, we implement a binary classification model for predicting traffic accident, achieving 0.7 F1 score on traffic accident dataset. Such model can be used to collect more traffic accident data.

1 Introduction

Numerous studies have focused on detecting traffic accidents [1], [2], [3]. However, limited research has been conducted on detecting accidents involving ego-vehicles (first-person perspective). Secondly, emerging architectures, such as advanced deep learning models and algorithms, have the potential to capture relevant information to perform computer vision task such as classification, detection, segmentation.

In this study, we aim to experiment various machine learning algorithm that can learn the differences between images by understanding the differences in their representations, and use such representation to classifies whether a video input contain traffic accidents scenario.

2 Related Work

2.1 Object Detection

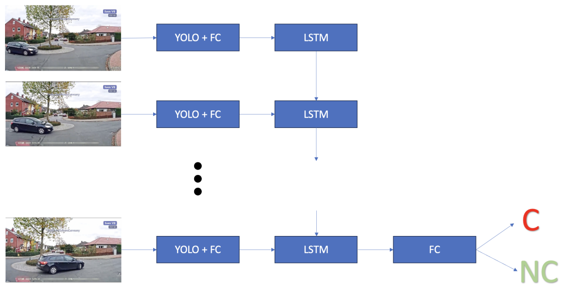
Object detection is crucial for obtaining information about various objects such as cars, traffic signs, and more. Features vector from the model, therefore, contain relevant representation for the task

- **YOLO Model**: The YOLO (You Only Look Once) model [4] is a state-of-the-art object detection model known for its speed and accuracy in detecting multiple objects within an image.

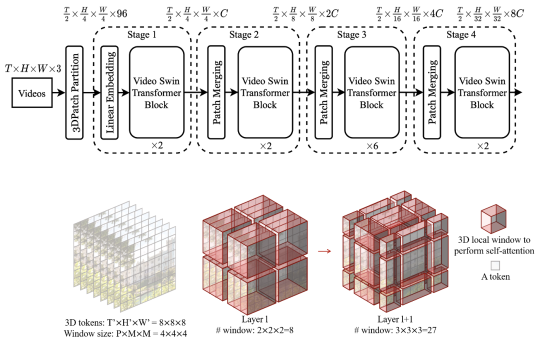
2.2 Video Classification

Video classification is important for capturing spatial and temporal information, which is necessary for collecting comprehensive data.

- **YOLOv1 with LSTM**: We believe the feature vectors extracted for multiple consecutive frames, together with LSTM [5], can capture both spatial and temporal information of the video input.
- **Video Swin Transformer**: Video Swin Transformers [6] model is another advanced model for video classification, utilizing transformer architecture to effectively handle spatial and temporal dependencies in video sequences.



(a) YOLOv1 with LSTM model architecture


$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times 96$
$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{T}{2}\times\frac{H}{8}\times\frac{W}{8}\times 2C$
$\frac{T}{2}\times\frac{H}{16}\times\frac{W}{16}\times 4C$
$\frac{T}{2}\times\frac{H}{32}\times\frac{W}{32}\times 8C$
$T\times H\times W\times 3$
Videos
3DPatch Partition
Stage 1
Linear Embedding
Video Swin Transformer Block
×2
Stage 2
Patch Merging
Video Swin Transformer Block
×2
Stage 3
Patch Merging
Video Swin Transformer Block
×6
Stage 4
Patch Merging
Video Swin Transformer Block
×2
3D tokens: T'×H'×W' = 8×8×8
Window size: P×M×M = 4×4×4
Layer l
# window: 2×2×2=8
Layer l+1
# window: 3×3×3=27
3D local window to perform self-attention
A token

(b) Swin 3D architecture

3 Approach

We experimented with 2 different architectures of YOLOv1 with LSTM. Specifically, we experimented how remove the last fully connected layer can affect the overall performance. YOLOv1 model get pretrained weight from an open-access GitHub repository [7] as the work is done on traffic dataset, and we reasoned the feature will be informative. The second experiment is for Video Swin Transformers with and without pretrained weights which been trained on Kinetics 400 dataset.

4 Experimental Result

4.1 Set up

Video Sampling: Divide videos into segments of approximately 4-5 seconds. We then randomly selected started point within 4-second segment and sample 10 consecutive frames. Labeling is defined as: Positive Labels: If the sequence of frames contains critical frames, label it as positive. Negative Labels: In all other cases, label the segment as negative.

- **Dashcam Video Dataset**: This dataset contains 50 videos. Critical scenarios are identified from manual labeled data, while all other frames are considered non-critical.

- **Car Crash Dataset [8]**: This dataset includes 800 videos, each lasting 5 seconds. Frames containing critical scenarios for the ego vehicle are randomly selected.

- **BDD100K Dataset [9]**: This dataset consists of 1000 videos, each 40 seconds long. Each video is divided into 4-second segments for analysis.

  YOLOv1 with LSTM is trained for 20 epochs, and Video Swin Transformer is trained with 16 epochs. All the model is trained with Adam optimizer and Cosine annealing schedule.

4.2 Metrics

Table 1: Model Performance

Model Name	F1 Score
Swin3D	0.178
YOLOv1 + LSTM	0.6
Swin3D with pretrained weights	**0.756**
YOLOv1 + LSTM without fc	0.601

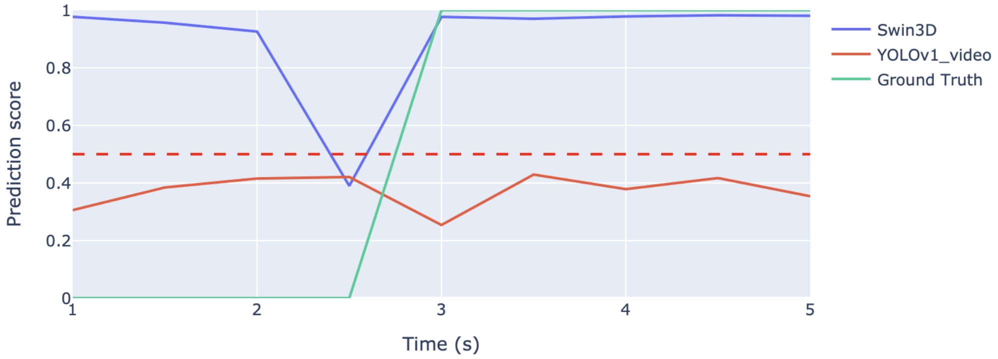


Figure 2: Critical classification prediction over time

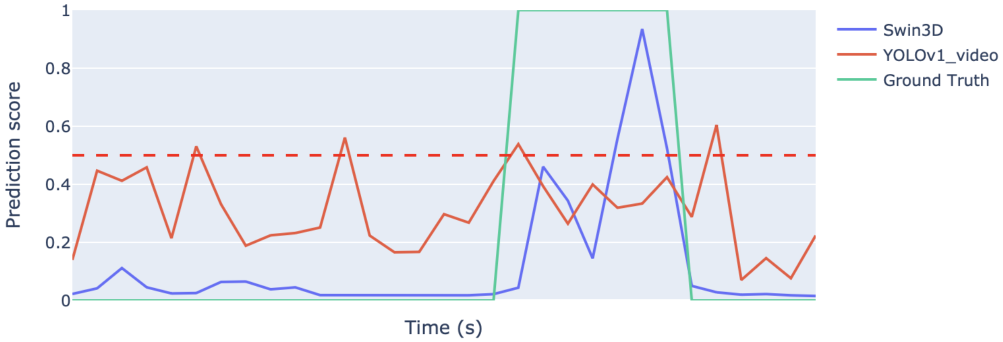


Figure 3: Critical classification prediction over time

5 Visualization

6 Conclusion and future work

We achieved reasonable performance on the task, with pretrained weights proving crucial for success. However, when using a YOLOv1 architecture combined with LSTM, the fully connected layer causes a loss of feature information, leading to poor generalization to the task.

In the future, we plan to focus on object tracking by applying our methods to extract the positions of objects of interest in a bird's-eye-view format. Our work can also support automated annotation tasks, enabling the creation of a diverse traffic accident dataset that could be instrumental in identifying safety concerns for autonomous vehicles.

Recent advancements in autonomous vehicles have led to significant improvements in various driving-related tasks, but ensuring the safety of these systems remains paramount. Current methodologies for testing the safety of autonomous driving systems encompass a wide range of techniques. However, there is still a need to generate realistic traffic accident scenarios to further enhance these efforts.

Inspired by Nvidia's work, which successfully generated trajectories using generative adversarial networks (GANs), we recognize certain limitations in the datasets used. Specifically, the initial distribution of information regarding normal traffic is collected from the NuScenes dataset, which could introduce bias and affect the development of autonomous driving systems in different countries. Our binary model addresses this issue by collecting traffic accident scenarios from various locations, thereby improving the initial distribution and enhancing the overall robustness of autonomous driving systems.

References

[1] Karishma Pawar and Vahida Attar. Deep learning based detection and localization of road accidents from traffic surveillance videos. *ICT Express*, 8(3):379–387, 2022.

[2] Hyeon-Cheol Son, Da-Seul Kim, and Sung-Young Kim. Vehicle-level traffic accident detection on vehicle-mounted camera based on cascade bi-lstm. *Journal of Advanced Information Technology and Convergence*, 10(2):167–175, 2020.

[3] Hadi Ghahremannezhad, Hang Shi, and Chengjun Liu. Real-time accident detection in traffic surveillance using deep learning. pages 1–6, 06 2022.

[4] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Technical report, CMU, 1997.

[6] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *CoRR*, abs/2106.13230, 2021.

[7] Alen Smajic. Real-time object detection for autonomous driving using deep learning. `https://github.com/alen-smajic/Real-time-Object-Detection-for-Autonomous-Driving-using-Deep-Learning`, 2020. Accessed: 2024-08-14.

[8] Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *ACM Multimedia Conference*, May 2020.

[9] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.