Accident Prediction: CHANAKYA FELLOWSHIP - iHub Anubhuti

PROJECT: Large Multi-Modal Foundation Model for Traffic Accident Analysis / Detection

Initial Work:

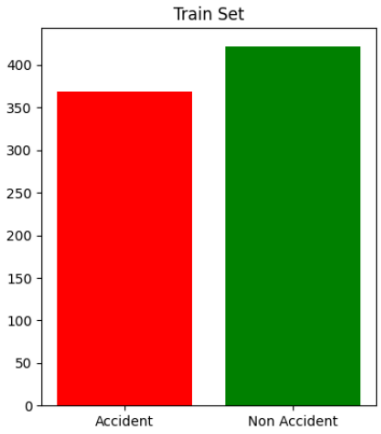
Problem Statement:

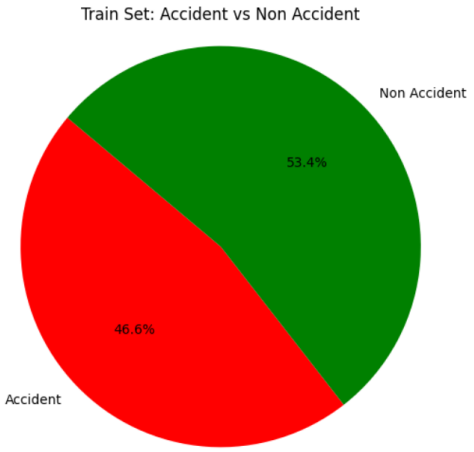
In this problem, we aim to create a detection system that can classify a scene from an image of vehicles on the road as an "**Accident**" or "**Non-accident**."

CCTV Image Dataset Description

Training Dataset:

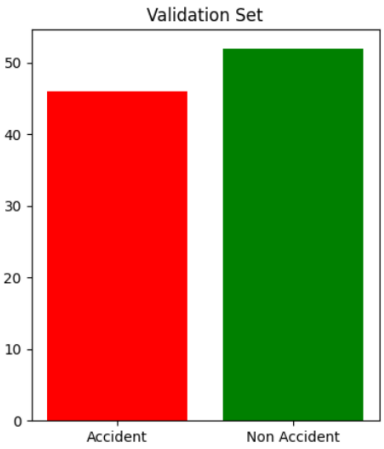
- **Total Accident Images**: 369
- **Total Non-accident Images**: 422
- **Total Images**: 791

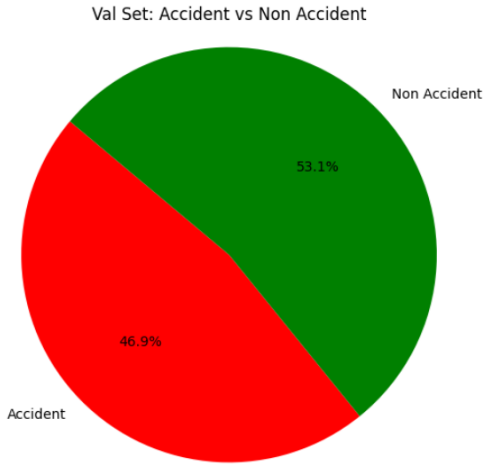




Validation Dataset:

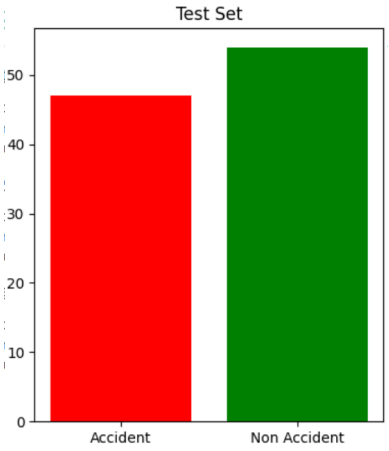
- **Total Accident Images**: 46
- **Total Non-accident Images**: 52
- **Total Images:** 98

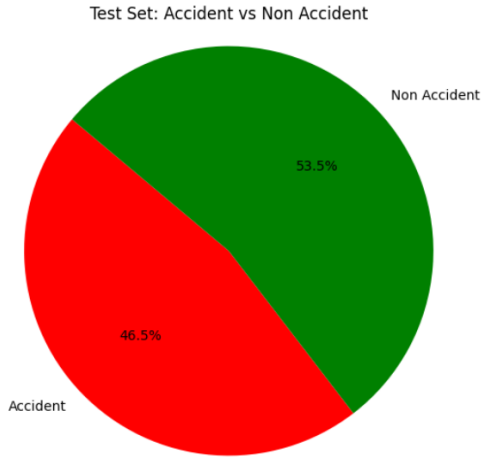




Testing Dataset:

- **Total Accident Images**: 47
- **Total Non-accident Images**: 54
- **Total Images**: 101





Sample Accident images:

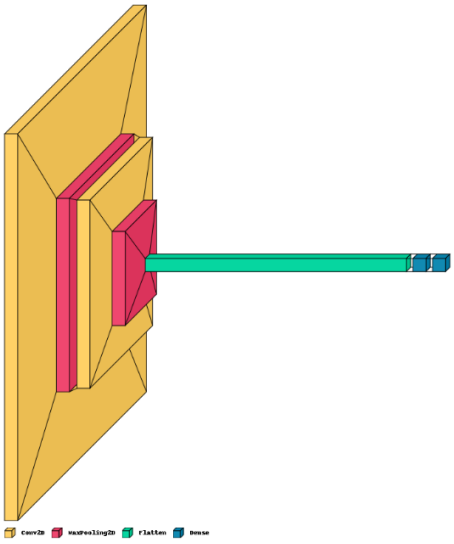



Sample Non-accident images:




MODEL ARCHITECTURE:

MODEL ARCHITECTURE:



```
model = Sequential([
    Conv2D(32, (3, 3), activation='relu',
    input_shape=(img_width, img_height, 3)),
    MaxPooling2D((2, 2)),
    Conv2D(64, (3, 3), activation='relu'),
    MaxPooling2D((2, 2)),
    Flatten(),
    Dense(128, activation='relu'),
    Dense(1, activation='sigmoid')
])
```

RESULTS:

Metric	Training	validation	Testing
Accuracy	**0.9209**	**0.9167**	**0.8899**
Precision	**0.9075**	**0.8953**	**0.9271**
Recall	**0.9198**	**0.9172**	**0.9163**
F1 score	**0.9136**	**0.9061**	**0.9217**

The best F1 score achieved on test data, which is **0.9217**, indicates that the model's performance in terms of both precision and recall is robust, maintaining a balance between accurately identifying positive cases (precision) and capturing all relevant positive cases (recall).

CURRENT WORK:

Literature Review

Introduction

With rapid advancements in artificial intelligence and multi-modal data processing, there is a growing interest in developing foundation models for traffic accident analysis and detection. Leveraging diverse data sources such as vehicular data, pedestrian behaviour, CCTV footage, weather conditions, and road infrastructure can yield a robust framework capable of uncovering the root causes of traffic accidents, thereby improving road safety and traffic management. This literature review highlights significant contributions in accident prediction, diagnostic frameworks, data integration techniques, and multi-modal traffic analysis.

Multi-Modal Data Integration for Traffic Accident Prediction

Integrating multiple data sources is critical for developing a robust accident detection model. Yang et al. (2016) pioneered a data-driven approach by creating an early warning system for traffic accidents, utilizing statistical analysis to detect risk factors and suggest preventive measures in Shanghai. Their model provides a scalable foundation that could be adapted for multi-modal integration in broader contexts (Yang et al., 2016). Similarly, Gang et al. (2016) introduced a diagnostic model based on rough set theory and decision tree algorithms, which captures and quantifies various traffic factors influencing accident morphologies. This approach provides a scientific framework for understanding accident risks, which is essential for developing a predictive model (Gang et al., 2016).

Advances in Traffic Accident Detection Techniques

Detecting accidents in real time and predicting traffic flow changes are essential for managing connected and automated transportation systems. Zhang et al. (2023) developed a grid-based parameter extraction technique combined with a support vector classification (SVC) model, achieving an impressive 87.72% detection rate in connected transport environments. The study emphasizes the importance of spatial and temporal data processing, which could greatly enhance multi-modal models for real-time detection (Zhang et al., 2023). Additionally, Pan & Wu (2017) implemented an SVM-based detection system using mobile sensors to identify urban traffic incidents, highlighting the complexities of urban

environments and the importance of variables like speed and lane changes in accurate detection (Pan & Wu, 2017).

GIS-Based Frameworks for Accident Analysis

Analyzing and visualizing accident-prone areas is crucial for preemptive traffic management and public safety planning. Ye et al. (2010) proposed a GIS-based model that identifies accident black spots and clusters these data points for visualization, supporting informed decision-making in urban planning (Ye et al., 2010). Furthermore, Sreedhar (2021) applied logistic regression models to identify conditions affecting accident severity and utilized heatmaps to visualize accident-prone areas. This work underlines the value of spatial data integration in revealing accident hotspots, which can be a vital feature in multi-modal models (Sreedhar, 2021).

Big Data and Traffic Accident Analysis Platforms

Leveraging big data technology, Hu & Zhao (2017) designed a platform for traffic accident analysis, emphasizing data processing and analytics. This system integrates various data sources, highlighting the potential of big data in analyzing and predicting traffic accidents on a large scale. Their work underscores the role of big data in handling diverse data sources within a multi-modal framework (Hu & Zhao, 2017).

Systems-Based Methods for Accident Analysis

Analyzing road accidents often involves understanding complex interactions between various contributing factors. Zhang et al. (2018) combined the HFACS (Human Factors Analysis and Classification System) and CFIM (Causal Factors Integration Model) frameworks to examine road safety, focusing on unsafe behaviors and environmental influences. This systems-based approach provides a comprehensive analysis of accident causation and supports the creation of multi-modal models that can detect patterns related to human factors (Zhang et al., 2018).

Summary

The reviewed studies underscore the potential of integrating multi-modal data for traffic accident analysis and detection. Advances in data mining, decision tree

models, GIS-based analysis, and big data processing contribute to a rich foundation upon which a comprehensive, multi-modal model can be built. These methods provide critical insights into accident risk factors, real-time detection techniques, and visual representations of accident-prone areas. Moving forward, integrating machine learning algorithms with these multi-modal data sources is expected to yield a sophisticated foundation model capable of enhancing road safety and supporting proactive traffic management.

Conclusion

Developing a large multi-modal foundation model for traffic accident analysis requires synthesizing insights from varied data processing techniques, machine learning models, and diagnostic frameworks. This model can effectively identify, predict, and mitigate traffic accidents by leveraging the advancements in real-time detection and spatial analysis. The foundation laid by prior studies provides a clear path forward for implementing a system capable of transforming traffic safety and accident prevention strategies.

References:

1. https://ieeexplore.ieee.org/document/8047095
2. https://www.tandfonline.com/doi/full/10.1080/03081060.2016.1231894
3. https://onlinelibrary.wiley.com/doi/10.1155/2023/5041509
4. https://ieeexplore.ieee.org/document/8104994
5. https://ieeexplore.ieee.org/document/5593800
6. https://www.ijraset.com/fileserve.php?FID=33280
7. https://link.springer.com/chapter/10.1007/978-3-319-67071-3_6
8. https://www.sciencedirect.com/science/article/abs/pii/S1369847816302820?via%3Dihub

Dataset Curation:

Required Datasets: We found some other datasets related to image modality but can't access them currently; we need help accessing them.

1. Traffic Accident Detection Video Dataset for AI-Driven Computer Vision Systems in Smart City Transportation (https://dx.doi.org/10.21227/tjtg-nz28) This dataset is at the IEEE-data port; thus, we need IEEE creds for accessing this dataset.
2. Anticipating Accidents in Dashcam Videos (https://aliensunmin.github.io/project/dashcam/) This requires filling out mentor credentials in *Google Forms* to access the dataset.

We currently extended our work to include video modality such that we can incorporate dashcam footage

We found a video dataset that has label-led dashcam videos of hugging faces. (
Hugging face dataset)

Current Dataset: Smart-Dashcam

Number of Videos: 136

Number of accident cases: 68

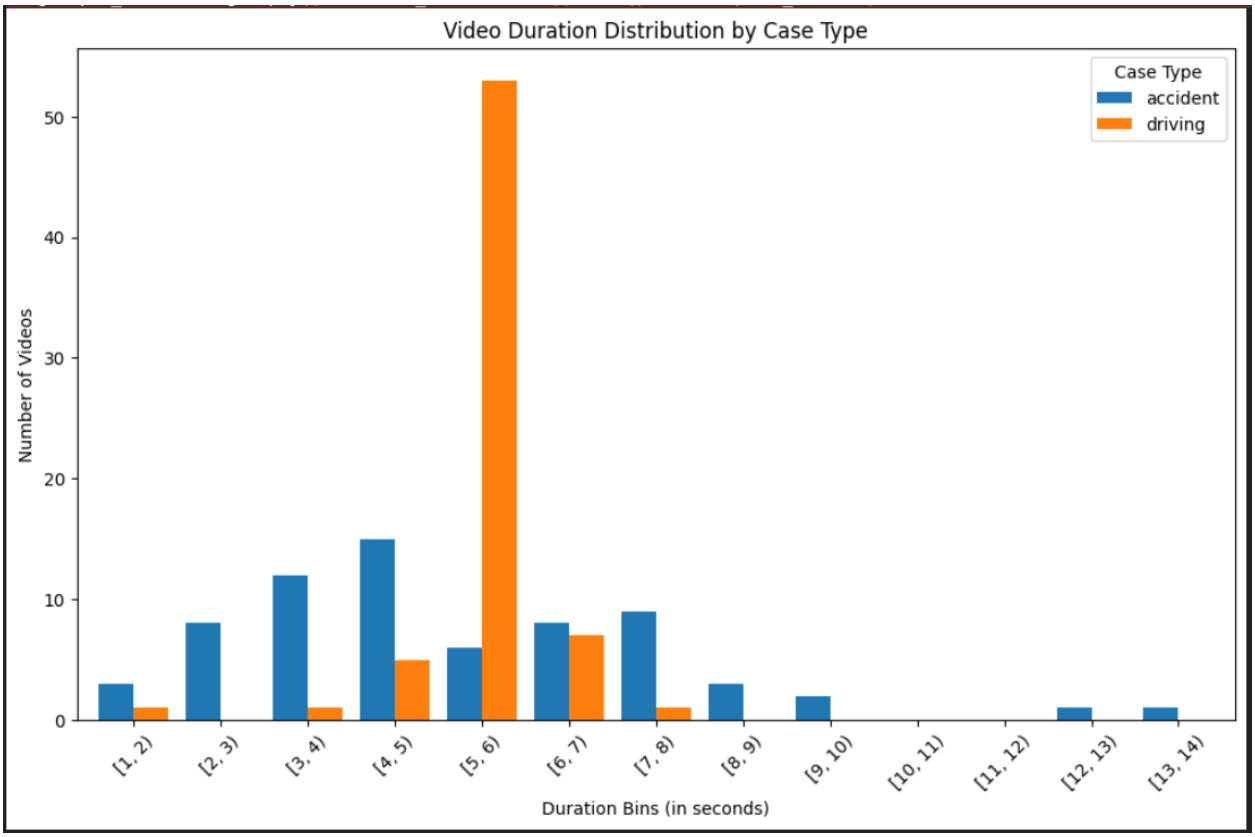
Number of driving cases: 68

Max duration of a video: 13.15 seconds

Minimum duration of a video: 1.47 seconds

Average duration: 5.41 seconds

Video Duration Distribution:



Data Preprocessing:

Video to Image Sequence

The frames are extracted from videos before the training. Utilizing every frame is impractical and inefficient, primarily due to resource constraints. Instead, a specified number of frames are extracted equally spaced throughout the video. For all the videos, the sequence length is
S (i.e. the number of frames considered from the video; Hyper parameter), and the `S = 40`.
Reasoning: This approach ensures that if an accident scenario exists within a segment and spans approximately the (length of the video in frames/sequence length) seconds, it will be present in the reduced subset of frames. This contrasts with a scenario where, for instance, only the beginning of a video is considered as a segment. In such cases, accident scenarios located towards the end may go unnoticed, or vice versa, leading to a potential oversight.

This allows us to process different-length videos, keeping the relationship between the frames of each video.

Multi-Image Resizing

The frames are then resized to a specific dimension of `C x H x W = 3 x 92 x 92`.
This standardization allows for handling multiple videos with varying dimensions and encoding schemes by resizing them to a consistent shape with the RGB channel of each image.

Resulting, the dataset now has n (136) data points.

Each data point : `40 x (3 x 92 x 92) :: S x (C x H x W)`

Label: 2 classes (Accident or Non-Accident)

The dataset is well balanced, with equal videos of both classes.
Moreover, the same S is used for each data point, removing the concern of larger videos having more frames.

Problem with Image Sampling Approach:

The approach of considering each image frame from a video as a separate data sample for accident detection presents significant challenges, especially when it comes to labelling these frames. In scenarios where we need to classify video sequences as either an "accident" or "driving," treating each frame as an isolated instance and trying to assign a label—such as "accident" or "driving"—to each image is inherently problematic

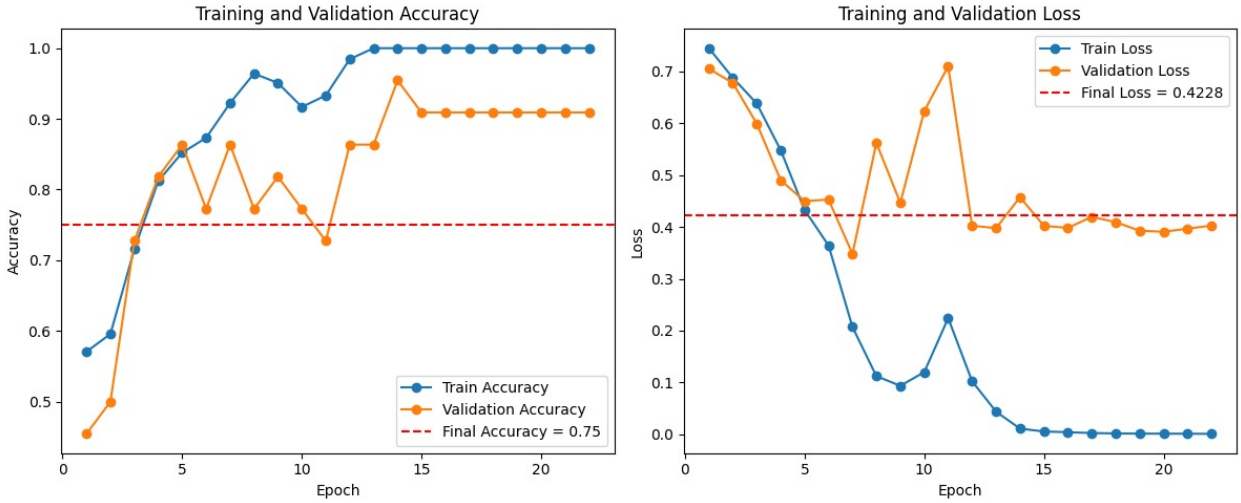
A single image extracted from a video lacks the temporal context required to understand the progression of events. Accident detection relies heavily on identifying a sequence of movements and changes that happen over time rather than a static, single frame. A frame taken from the beginning of an "accident" video may show a perfectly normal driving situation, with no signs that an accident is about to occur. Consequently, labelling such an early frame as an

"accident" would be misleading and inaccurate. This demonstrates a critical flaw in single-frame labelling. Without a broader temporal context, it's impossible to accurately determine if an isolated frame is part of an accident or regular driving.

If all frames from a video labelled as "accident" are uniformly assigned the label "accident," the model may develop a skewed understanding of what constitutes an accident. For instance:

- The model could misinterpret normal driving frames as part of an accident scenario, reducing its ability to distinguish between safe driving and accidents.
- The lack of distinction between the buildup to the accident and the accident itself might lead to poor model performance, as it’s forced to learn from incorrectly labelled data.

We tried using 3D Convolutional kernel (2 Layers) followed by mlp layers.
This didn't result in good accuracy at the test set (Plateau to 75% only).



The Need for a Sequential or Temporal Approach (CNN LSTM):

Model Architecture

In this baseline model, each video segment is processed as a single data sample through a CNN-LSTM architecture.

The CNN first extracts spatial features from each frame using a convolutional layer, then max pooling to condense important details and dropout to prevent overfitting. This set of layers is repeated three times with varying parameters, enabling the model to capture increasingly complex features.

After CNN processing, the output is flattened and passed to an LSTM layer.

LSTM examines the sequence of frame-level features to detect temporal patterns.

Thus, CNN-based spatial feature extraction (parallelizable) and LSTM is temporal feature on spacial features.

Usefulness for Sequential Accident Detection

This CNN-LSTM approach effectively identifies accident scenarios by analyzing spatial and temporal features within a single data sample (the full video segment). The CNN extracts critical frame-specific information, while the LSTM processes these frames sequentially, capturing the progression from normal driving to potential accident indicators, like sudden manoeuvres or changes in object proximity. Treating the video as a single data sample allows the model to assess the entire event sequence cohesively, improving its ability to detect gradual transitions and reliably identify accident scenarios.

NOTE: We are currently refining the CNN-LSTM architecture by experimenting with various hyperparameters, such as the optimal number of frames per video segment, to improve model accuracy and reliability. Additionally, we are exploring a wider range of diverse datasets relevant to accident detection, with several reference datasets already identified to support this effort. These adjustments aim to enhance the model's ability to capture nuanced spatial and temporal patterns essential for detecting accident scenarios effectively.