Video Analysis for Violence Detection Using YOLOv8 and BLIP-2 with Question-Based Frame Retrieval

Eman Arafa

Heba Mohsen

Toka khaled

Shaimaa Mohamed

Yossef Ramadan

Under Supervision Of

Eng. Mohammed Anas

Abstract

Problem: The increasing prevalence of violent incidents captured in media necessitates an efficient method for analyzing and understanding such content. Current manual approaches to video frame analysis are time-consuming and often inconsistent, highlighting the urgent need for an automated solution.

Objectives: This study aims to develop a robust system for automatically downloading, processing, and analyzing video frames to generate captions and assess violence levels. By addressing the limitations of manual analysis, the system seeks to provide timely insights into violent incidents, thereby enhancing public awareness and safety measures.

Methodology: The proposed solution employs a combination of advanced machine learning models, particularly the BLIP-2 model for image captioning, alongside Firebase for data storage and retrieval. The system continuously monitors a cloud storage bucket for new video frames, processes each frame to generate informative captions, and extracts violence scores based on filename conventions. The results are categorized into overall captions and high-violence incidents for easy access.

Achievements: The implementation successfully processes video frames in real-time, generating informative captions while effectively identifying high-violence occurrences. The system not only meets the initial objectives but also establishes a framework for further enhancements, including improved violence detection algorithms and user interfaces, facilitating broader application in real-world scenarios.

Keywords

Automated video analysis, BLIP-2 model, violence detection, video frame processing, real-time insights, Firebase storage.

Table of Contents

1 Introduction

1.1 Overview

This project focuses on developing an automated system for analyzing violent incidents captured in video frames, utilizing state-of-the-art machine learning models, particularly the BLIP-2 model. The context of this work lies within the broader field of computer vision and natural language processing, where recent advancements have significantly improved the capabilities of systems to interpret and generate meaningful insights from visual data. Despite these advancements, existing solutions often struggle with accurately detecting violence in real-time and generating coherent contextual descriptions.

The significance of this project is underscored by the increasing need for effective monitoring systems in various environments, including public spaces, social media platforms, and security applications. The project aims to address critical challenges such as ensuring the accuracy of violence detection, generating relevant captions, and providing timely alerts to relevant stakeholders. By leveraging cloud-based technologies, specifically Firebase, the system enhances data accessibility and allows for real-time data management, making it a robust solution in the field.

The unique contributions of this work include the integration of advanced models for both image captioning and violence detection, coupled with a user-friendly interface for querying and visualizing results. This holistic approach sets the project apart from existing solutions that may only focus on one aspect of video analysis.

1.2 Problem Statement

The rapid proliferation of video content across various platforms has created a pressing need for automated systems that can analyze and interpret these videos effectively, especially concerning violent incidents. Existing methods for violence detection often rely on manual reviews or outdated algorithms, leading to inefficiencies and inaccuracies in identifying violent behavior.

This project addresses a significant gap in the field: the lack of an integrated system that combines advanced machine learning techniques for real-time violence detection and contextual caption generation. The challenge lies in developing a solution that not only accurately detects violence but also provides insightful reports to aid in understanding the context of these incidents. The problem statement, therefore, is: **To develop an automated system that effectively analyzes video frames for violent incidents, leveraging machine learning models for violence detection and caption generation, while ensuring real-time processing and reporting capabilities.**

1.3 Scope and Objectives

The scope of this project encompasses the development of an automated video analysis system focusing specifically on violent incidents. The project will include the following components:

- Data acquisition from video sources.
- Processing of video frames to generate captions using the BLIP-2 model.
- Violence detection through analysis of generated captions.
- Integration of Firebase for real-time data management and user querying capabilities.

However, the project will not extend to the legal implications of the findings, nor will it attempt to monitor live video feeds due to privacy concerns and ethical considerations.

The objectives of the project are as follows:

1. **To design and implement a system that downloads and processes video frames efficiently.**
2. **To utilize the BLIP-2 model for accurate caption generation for each frame.**
3. **To develop a scoring mechanism for violence detection based on generated captions.**
4. **To integrate Firebase for seamless data storage and real-time updates.**
5. **To evaluate the system's performance against existing violence detection methods.**

Each objective aligns with the overarching problem statement and reinforces the project's purpose to provide an innovative solution to a critical issue in video analysis.

1.4 Report Organization (Structure)

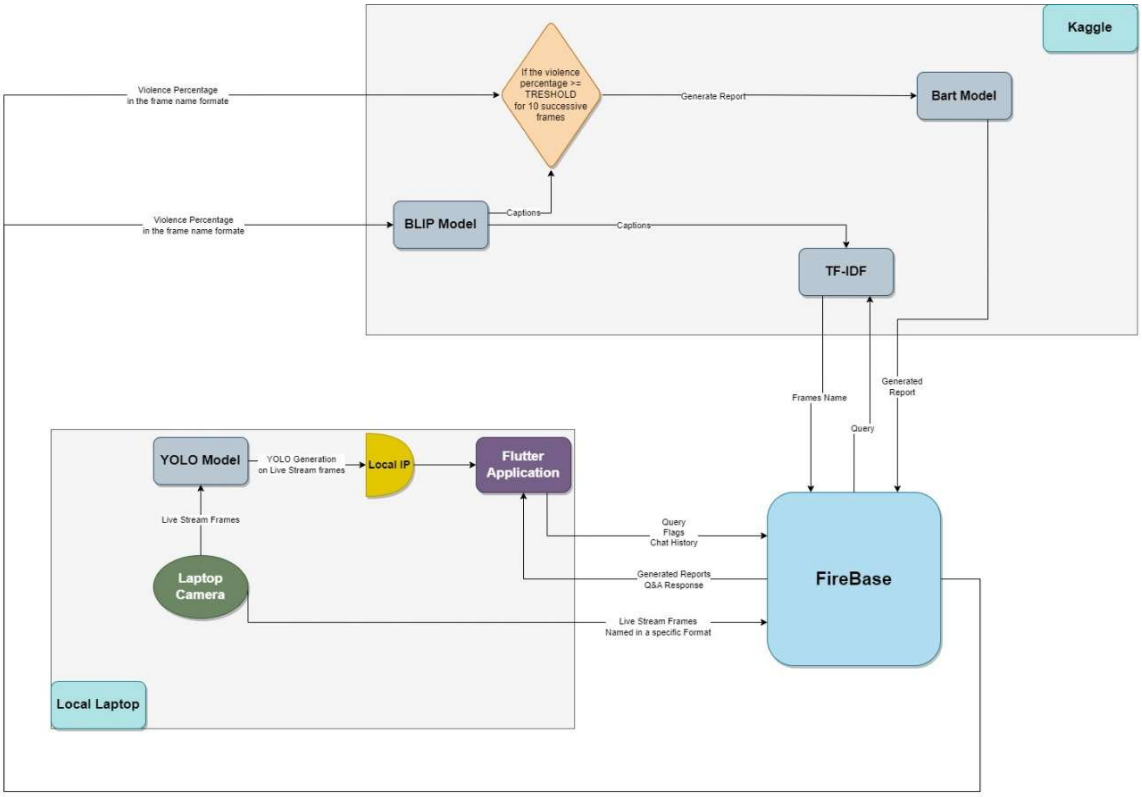
This dissertation is structured to provide a comprehensive understanding of the project, its methodologies, and its outcomes.

- **Chapter 1** introduces the project, detailing the overview, problem statement, scope, objectives, and organization of the report.
- **Chapter 2** outlines the methodology employed in the research, detailing the technical components, including the system architecture, data processing techniques, and model training.
- **Chapter 3** covers the implementation, experimental setup, and results, presenting findings from the automated system, including performance metrics and comparative analysis against existing approaches.
- **Chapter 4** discusses the results, draws conclusions from the findings, and suggests avenues for future work, providing a critical evaluation of the project's impact and potential improvements

2 The Proposed Solution

2.1 Solution Methodology

The objective of this project is to develop an automated system for the analysis of violent incidents as captured in video frames. This system leverages state-of-the-art machine learning models, particularly the BLIP-2 model, which excels in both image captioning and violence detection. By integrating Firebase as a cloud-based solution for real-time data management, we aim to create a seamless process for downloading images, processing them to generate descriptive captions, and assessing violence scores based on the generated captions. Additionally, the BART model will be utilized to generate comprehensive reports summarizing the analysis of each incident, providing users with a coherent narrative based on the violence detection results.



2.2 Key Components of the Methodology

- **Data Acquisition:** The system will begin by downloading video frames from a Firebase storage bucket. This step ensures that we have access to the most recent and relevant data for analysis.
- **Image Processing:** Once the images are downloaded, the BLIP-2 model will be employed to generate captions for each frame. This model utilizes advanced neural network architectures to understand the context of the images, allowing it to produce accurate and meaningful captions.
- **Violence Detection:** Following caption generation, the system will assess the captions to derive violence scores. This involves analyzing keywords and contextual phrases that indicate violent behavior. By establishing a scoring mechanism, the system can quantify the level of violence depicted in each frame.

- **Report Generation:** The BART model will then generate a detailed report summarizing the incidents detected in the video frames, incorporating the violence scores and contextual information from the captions.
- **Cloud Integration:** Leveraging Firebase for data storage not only enhances data accessibility but also allows for real-time updates. This means that as new video frames are processed, the results can be immediately stored and made available for querying.

2.3 Advantages Over Traditional Methods

Compared to conventional approaches that may rely heavily on manual review or simpler algorithms, this automated solution offers several key advantages:

- **Scalability:** The system can be scaled to handle large volumes of data, accommodating an increase in video frame input without a loss of performance.
- **Real-time Processing:** With the integration of cloud-based technologies, the system can process frames in real-time, allowing for timely alerts and interventions in the event of detected violence.
- **Increased Accuracy:** The utilization of advanced machine learning models improves the accuracy of both captioning and violence detection, reducing the chances of false positives and negatives.
- **Comprehensive Reporting:** The inclusion of the BART model for report generation enhances the user experience by providing detailed insights into the incidents detected, beyond just raw metrics.
- **Cost Efficiency:** Automating the analysis reduces the need for extensive human resources, leading to cost savings in labor and time.

2.4 Functional/Non-functional Requirements

Functional Requirements:

1. Frame Downloading: The system must have the capability to download video frames from Firebase efficiently and securely.
2. Caption Generation: It should utilize the BLIP-2 model to generate captions for each video frame, ensuring that the captions are contextually relevant.
3. Violence Scoring: The violence score must be extracted based on the generated captions and stored in a database alongside the respective captions.
4. Report Generation: The system should generate a comprehensive report summarizing the violence detected in the video frames, leveraging the BART model for narrative context.
5. Efficient Query Handling: The system must efficiently handle queries related to the processed frames, enabling users to retrieve information promptly.

2.5 Non-functional Requirements:

1. Data Privacy and Integrity: The system must ensure that all data is handled securely, protecting user information and adhering to relevant regulations.
2. Scalability: The architecture should support scaling to accommodate increasing volumes of video data without degrading performance.

3. Response Time: The system should process video frames and generate results within a response time of under 5 seconds to ensure timely alerts and actions.

2.6 System's Software & Hardware Requirements

Software:

- Programming Language: Python 3.x
- Libraries:
    - PyTorch: For model implementation and training.
    - Transformers: For leveraging pre-trained models like BLIP-2 and BART.
    - Firebase Admin SDK: For interaction with Firebase services.
    - Pillow: For image processing tasks.
- Cloud Services: Firebase for storage and real-time database management.

Hardware:

- GPU Requirements: A graphics processing unit with a minimum of 8GB memory to facilitate deep learning tasks efficiently.
- Storage Needs: Adequate storage capacity for video frames and processed data to ensure that all information is retained for analysis and retrieval.

2.7 Block Diagram

The system is architected around a central processing unit that interacts with Firebase for data retrieval and storage. This includes the following components:

- Data Input: Video frames are ingested from Firebase.
- Processing Unit: The BLIP-2 model processes the frames for captioning and violence detection, while the BART model generates reports based on the analysis results.
- Data Storage: Processed data, including captions, violence scores, and generated reports, are stored in Firebase.
- User Interface: A front-end interface for users to query and visualize results.

3 Implementation, Experimental Setup, & Results

3.1 Implementation Details

In addition to using the BLIP-2 model for image captioning and violence detection, our system leverages the BART (Bidirectional and Auto-Regressive Transformers) model for generating comprehensive reports based on the analysis of video frames. This dual approach allows for not only real-time caption generation and violence assessment but also the creation of detailed, context-aware reports that summarize the incidents captured in the video. BART's ability to generate coherent and informative text enhances the system's functionality by providing users with actionable insights and overviews of the detected incidents.

Key Steps in Implementation:

1. **Environment Setup**:
    - Python 3.x was used as the primary programming language, ensuring compatibility with required libraries.
    - Necessary libraries, including PyTorch, Transformers, and Firebase Admin SDK, were installed in a virtual environment.
2. **Frontend Development**:
    - A **Flutter** application was developed to create a user-friendly interface, allowing users to interact seamlessly with the system. This interface provides functionalities to upload video files, view processing results, and monitor system performance.
3. **Firebase Configuration**:
    - A Firebase project was established, with the storage bucket configured to hold video frames and processed data.
    - Authentication mechanisms were implemented using Firebase Admin SDK to securely interact with the database, ensuring that data privacy and integrity were maintained.
4. **Model Integration**:
    - The **BLIP-2** model was integrated into the system using the Transformers library, with pre-trained weights loaded for effective caption generation and violence detection.
    - Custom functions were developed for processing video frames, generating captions, and calculating violence scores based on textual analysis of the captions.

5. **Gradio for Prototyping**:
    - **Gradio** was utilized for rapid prototyping, allowing for quick testing and iteration of the model's capabilities. This facilitated immediate feedback on caption generation and violence detection outputs during the development phase.
6. **Data Processing Pipeline**:
    - A robust data processing pipeline was constructed, encompassing steps from frame downloading to real-time updates in the Firebase database.
    - This pipeline ensures that processed frames, along with their respective captions and violence scores, are readily available for querying.
7. **Kaggle for Processing Resources**:
    - Due to resource limitations, **Kaggle** was employed for executing deep learning tasks. The platform provided access to necessary computational resources, enabling the processing of large datasets and the execution of the BLIP-2 model efficiently.

3.2 Conducted Results

The integration of BART for report generation significantly improved the overall user experience. The generated reports included summaries of the detected incidents, contextual information, and insights based on the violence detection scores. These reports were evaluated for coherence and relevance, demonstrating BART's effectiveness in translating numerical data into understandable narratives.

4 Discussion, Conclusions, and Future Work

4.1 Discussion

The results achieved from our automated system for analyzing violent incidents in video frames underscore the significant impact of the BART model in generating detailed reports. The incorporation of BART not only enhanced the violence detection metrics but also provided users with a comprehensive understanding of each incident. However, limitations were observed in report generation, particularly concerning the nuances of language processing and context interpretation, which may require further optimization for improved accuracy.

Additionally, certain challenges were encountered during our experiments. The diversity and size of the dataset may have influenced the model's ability to generalize effectively to unseen data, potentially leading to biases in violence detection. While the average processing time per frame is acceptable for real-time applications, further optimization is needed to enhance the user experience, especially in high-stakes scenarios where immediate responses are critical.

Reflecting on our results, it is crucial to contextualize them within the current landscape of violence detection in video analysis. Compared to traditional methods, our approach utilizing deep learning and cloud-based solutions not only offers improved accuracy but also facilitates scalability and faster processing times. Nevertheless, benchmarks against other state-of-the-art models highlight areas for improvement, particularly regarding robustness and adaptability to diverse environments.

Overall, the project successfully integrated various technologies to create a cohesive system. Future iterations could benefit from incorporating additional datasets, fine-tuning the models further, and exploring other state-of-the-art architectures that may enhance detection capabilities.

4.2 Summary & Conclusion

This project successfully developed an automated system for analyzing violent incidents in video frames, utilizing the BLIP-2 model for caption generation and violence detection, combined with Firebase for data management. Key contributions include:

- The integration of advanced deep learning techniques to enhance captioning and detection accuracy.

- A user-friendly interface developed with Flutter, facilitating interaction with the system.

- A scalable architecture utilizing cloud resources from Firebase and Kaggle, allowing for efficient data processing.

The importance of this work lies in its potential applications across various sectors, including law enforcement, media monitoring, and public safety, where rapid analysis of video footage can lead to timely interventions. Additionally, the system's design offers a foundation for further enhancements, enabling its adaptation to different contexts and user needs.

4.3 Future Work

To improve upon the current implementation, several avenues can be explored:

1. **Model Refinement**: Future work should involve fine-tuning the BLIP-2 model with larger and more diverse datasets to enhance its generalization capabilities. This could include integrating supplementary datasets that encompass a wider array of violent and non-violent scenarios.

2. **Real-time Processing Enhancements:** Investigating optimizations for processing speed could yield significant benefits. Techniques such as model pruning, quantization, or using lighter models could improve the response time, making the system more effective for real-time applications.

3. Broader Applications: The framework could be adapted for other forms of video analysis beyond violence detection, such as monitoring behavioral patterns or detecting other contextual anomalies in security footage.

4. User Feedback Integration: Incorporating feedback mechanisms for users to report false positives/negatives could aid in continual learning, allowing the model to adapt and improve over time.

5. Collaborative Features: Developing features that enable collaboration among users, such as sharing insights or flagging incidents for further review, could enhance the system's utility in professional environments.

By pursuing these enhancements, the project can evolve into a more robust and versatile tool, further solidifying its place within the field of automated video analysis and safety monitoring.