Crime detection and classification based on Deep learning technique

By:

- **Seham Ahmed Mohamed**

Under Supervision of :

- **Prof. Dr. Kareem Ahmed**
- **Prof. Dr. Mohamed Ali Saleh**
- **Dr. Doaa Shibl**

Overview

¨Introduction

¨Problem Statement.

¨Motivation And Objectives.

¨Related Work.

¨Introduction To Deep Learning

¨Proposed Model.

¨Reference

Introduction

In recent years, the use of surveillance cameras has become increasingly prevalent in public spaces, such as shopping malls, airports, and city streets. These cameras serve as important tools for monitoring and ensuring public safety. However, the sheer volume of video footage generated by these cameras makes it challenging for human operators to effectively analyze and respond to potential threats in real-time.

Introduction Cont’d

Real-World Violence Detection in Surveillance Cameras aims to address this challenge by leveraging advances in machine learning and deep learning techniques. The goal is to develop intelligent systems that can automatically detect and recognize violent events or behaviors from live video feeds, alerting security personnel to take appropriate actions swiftly.

8321 POLICIA NACIONAL
8315 EVALUADOR VISUAL DE EMERGENCIA
8316 EVALUADOR VISUAL DE EMERGENCIA
8309
8304
8322

Problem Statement

- The problem is that existing surveillance systems heavily rely on human intervention to detect and report violent activities, which can lead to delayed responses and missed incidents. Therefore, there is a critical need to develop intelligent systems that can automatically detect and recognize real-world violence in surveillance camera feeds.

Motivation And Objectives

- The primary motivation is to create safer environments for individuals and communities by identifying and intervening in violent situations promptly and automated violence detection systems can contribute to the collection of evidence for legal proceedings and post-incident analysis. Recorded footage containing violent incidents can be used to aid investigations, identify suspects, provide objective evidence, and reconstruct events accurately.

Related Work

Dataset	Research	Year	Work	RESULTS(METRICS)
UCF-CRIME	Virender Singh and Swati Singh, Real-Time Anomaly Recognition Through CCTV Using Neural Networks	2020	They proposed a neural network used is a convolutional network, specifically the inceptionV3 model by Google, which is pre-trained on ImageNet. This model is used for feature extraction from images and simplifies the input for the second neural network. Transfer learning is applied to leverage the knowledge gained from training on ImageNet and reduce the training time. The second neural network is a recurrent neural network (RNN) that analyzes the sequence of actions in a given time period in videos. Its purpose is to classify video segments as either threatening or safe.	the overall accuracy of the model is 82.23% with reduced overfitting
UCF-CRIME	Soheil Vosta and Kin-Choong Yow, A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras	2022	The researchers constructed their model using Residual Networks (ResNets) for efficient feature extraction in deep neural networks. In the next phase, they employed Convolutional LSTM (ConvLSTM) as a recurrent network to detect anomalies in their video dataset. The approach involves dividing each video file into sequences of frames, with the input to the Convolutional Neural Network (CNN) being the difference between each frame and the next one. The output of the CNN (ResNet50) is then passed to the ConvLSTM. This process is repeated for all frames, and the output is further processed through a max-pooling layer and multiple fully connected layers to obtain the desired outcome.	81.71% in AUC

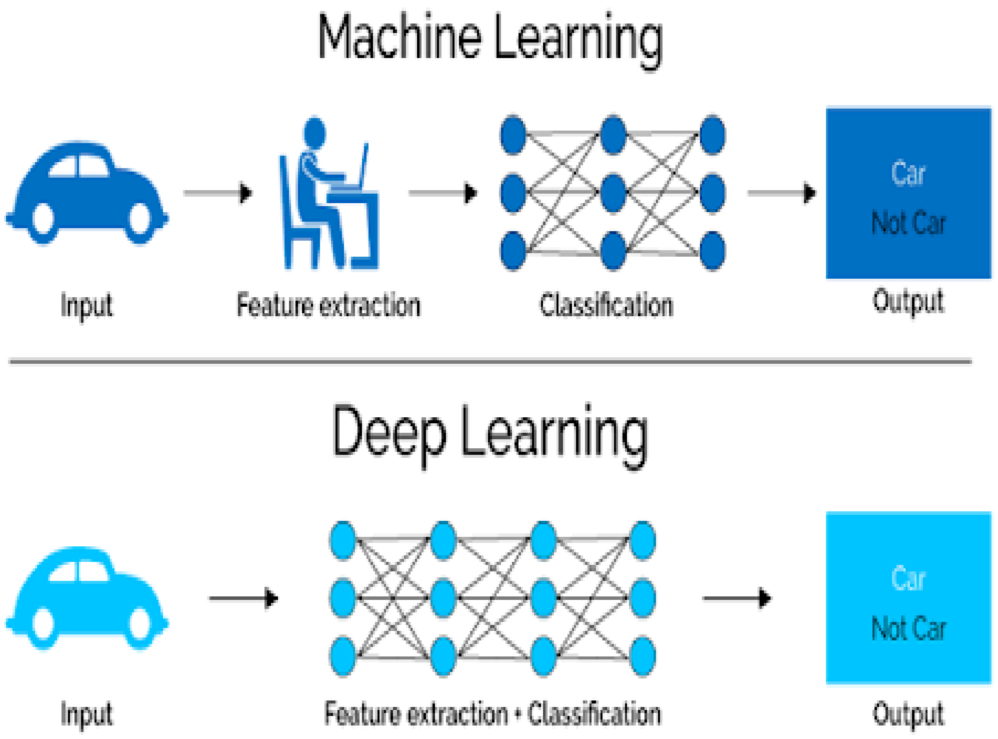
Related Work Cont’d

Dataset	Research	Year	Work	RESULTS(METRICS)
UCF-CRIME	Jhih-Ciang Wu and He-Yen Hsieh, Self-Supervised Sparse Representation for Video Anomaly Detection	2022	The researchers developed a framework called S3R (Self-Supervised Sparse Representation) to address two tasks: oVAD (online voice activity detection) and wVAD (wake-word voice activity detection). S3R utilizes a dictionary and self-supervised learning to model feature-level anomalies. It consists of two modules: en-Normal, which reconstructs normal-event features, and de-Normal, which filters out normal-event features. The researchers employed self-supervised techniques to generate pseudo anomaly/normal data based on the learned dictionary. This data is then used to guide the training of their anomaly detector.	80.47 in AUC
UCF-CRIME	Yingxian Chen and Zhengzhe Liu2, Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection	2022	The researchers introduced a framework called MGFN (Magnitude Guided Focus Network) for anomaly detection. MGFN incorporates a Glance-and-Focus module and a Magnitude Contrastive loss. The Glance and Focus mechanism mimics the human visual system by combining global context and local features effectively. They also proposed a Feature Amplification Mechanism (FAM) to improve the model's understanding of feature magnitudes. To learn a scene-specific distribution of feature magnitudes across videos, they introduced a Magnitude Contrastive loss, which encourages a clear separation between normal and abnormal feature magnitudes.	I3D-RGB X In AUC 86.98 VideoSwin-RGB X in AUC 86.67

Related Work Cont’d

Dataset	Research	Year	Work	RESULTS(METRICS)
UCF CRIME	Aswathy K. Cherian and and E. Poovammal Anomaly Detection in Real-Time Surveillance Videos Using Deep Learning	2021	They proposed a deep MIL framework where each video is treated as set and small segments of the video are treated as instance of these videos and I3D algorithm and a three layered FC neural network Feature extraction and classification	81% In AUC Accuracy I3D 25.9 C3D 23.0
UCF CRIME	Waqas Sultani and Chen Chen, Real-world Anomaly Detection in Surveillance Videos	2019	The researchers proposed a deep learning approach for detecting anomalies in surveillance videos. They acknowledged that using only normal data may not be sufficient for detecting complex real-world anomalies. To address this, they aimed to leverage both normal and anomalous videos. Instead of manually annotating anomalous segments in training videos, they employed a deep Multiple Instance Learning (MIL) framework with weakly labeled data to learn a general model for anomaly detection. To evaluate their approach, they also introduced a large-scale anomaly dataset containing a diverse range of real-world anomalies.	75.41% In AUC

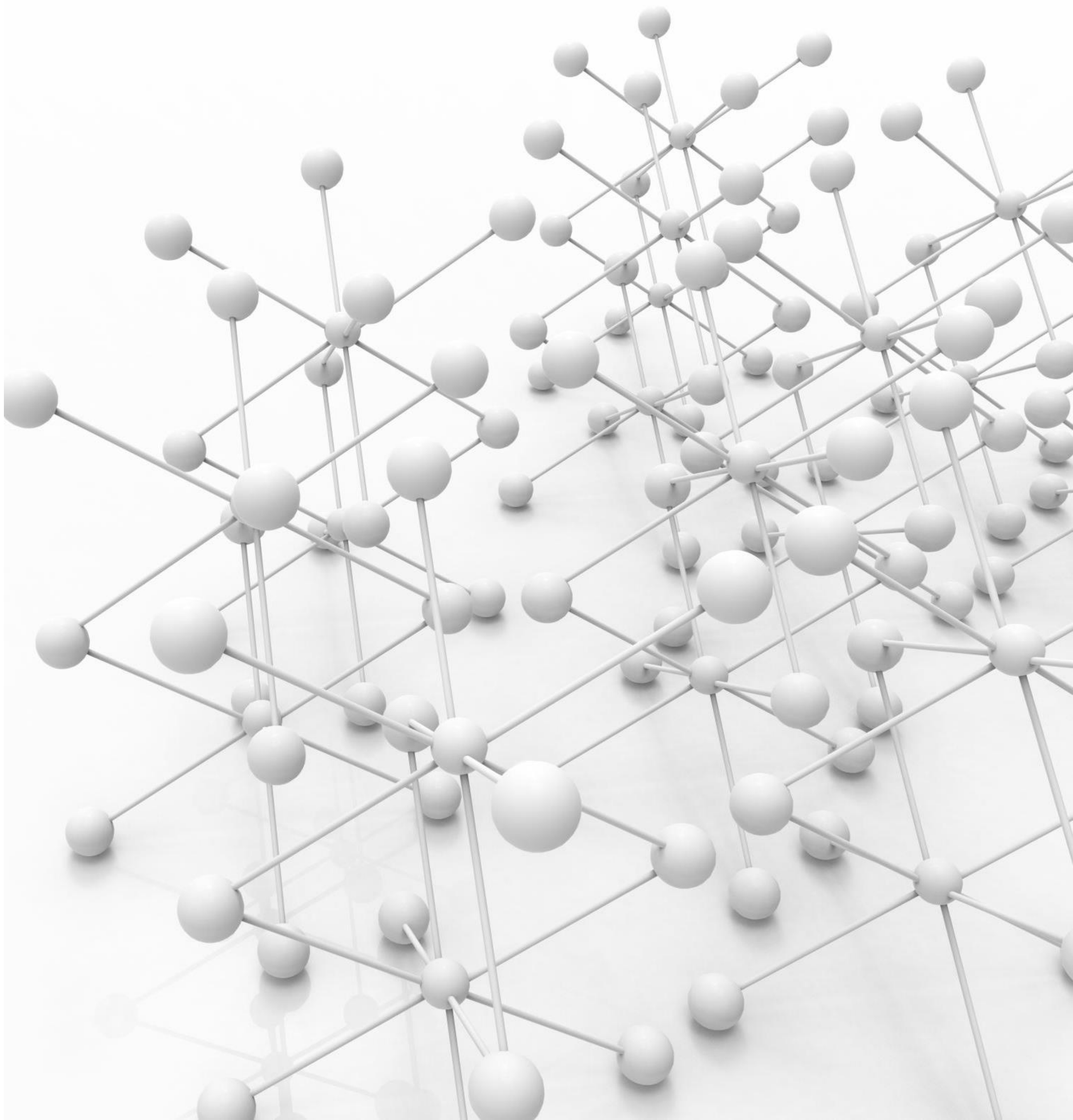
Introduction To Deep Learning



- **A machine learning** focuses on the development of algorithms that can learn patterns and make predictions or decisions based on data.
- **Deep learning** is a subset of machine learning that focuses on using artificial neural networks with multiple layers to learn hierarchical representations of data. If you provide the system **tons of information,** it begins to understand it and respond in useful ways.

Deep Learning Architectures

1. ResNets
2. VGG Net
3. NasNet
4. Densenet



ResNets

- **ResNet, short for Residual Neural Network, is a deep learning architecture that addresses the challenges of training deep neural networks by utilizing residual connections. It enables the training of extremely deep networks while maintaining good performance, making it a powerful tool for image classification and other computer vision tasks.**
- **ResNet, or Residual Neural Network, offers several advantages that have contributed to its popularity and success in various computer vision tasks:**

1. **Facilitates Training of Deeper Networks**
2. **Improved Gradient Flow**
3. **Enhanced Accuracy**

VGG Net

- GG Net, or the Visual Geometry Group Network, is a convolutional neural network architecture that has made significant contributions to image classification tasks. It was developed by the Visual Geometry Group at the University of Oxford.
- VGG Net is characterized by its simplicity and uniformity in design. It consists of several convolutional layers followed by max-pooling layers, with fully connected layers at the end for classification.
- The main contribution of VGG Net is its simplicity and uniformity in architecture. The network consists of several stacked convolutional layers, followed by max-pooling layers to reduce spatial dimensions, and finally, fully connected layers for classification. VGG Net demonstrated that increasing the depth of the network significantly improves performance on image classification tasks.
- Common features:

1. Simplicity and Uniformity
2. Flexible Architecture

NasNet

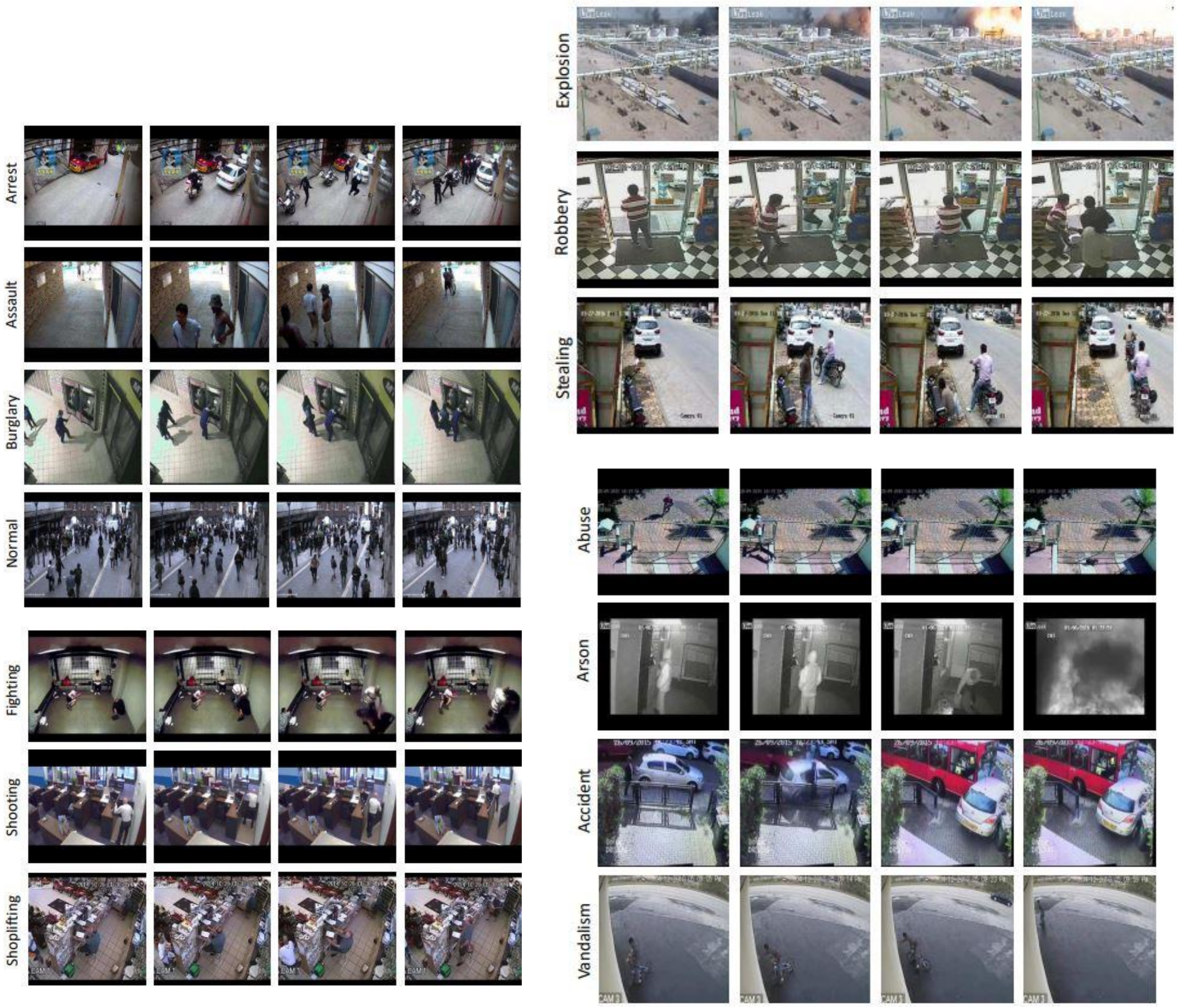
- NasNet, short for Neural Architecture Search Network, is an architecture developed using neural architecture search (NAS) techniques. NAS is a method that automates the design of deep learning architectures by allowing the network to search and discover the optimal architecture for a given task.
- NasNet utilizes reinforcement learning algorithms to search and learn the architecture that best fits the data and task at hand. Instead of manually designing and tuning architectures, NasNet automatically explores a large search space of possible architectures, optimizing for performance metrics such as accuracy or efficiency.
- In addition to its high performance advantages, NasNet also offers practical benefits such as reduced human effort in architecture design and faster development cycles. By automating the architecture search process, NasNet significantly reduces the need for manual trial-and-error exploration of different architectures.

Densenet

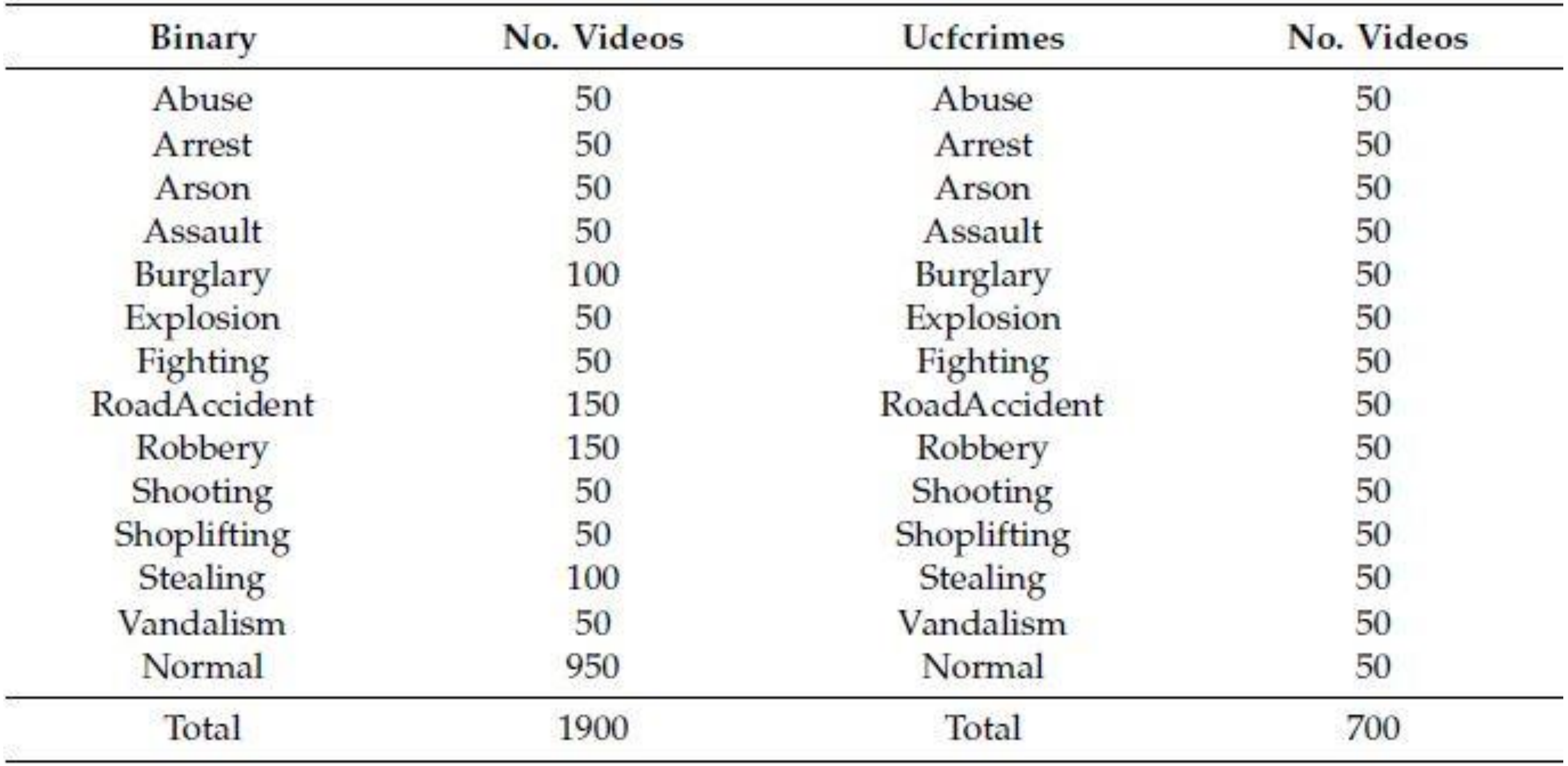
- DenseNet, short for Dense Convolutional Network, is a deep learning architecture that has gained popularity in the field of computer vision. It addresses the challenges of information flow and gradient propagation in deep networks by introducing dense connections.
- In DenseNet, each layer is connected to every other layer in a feed-forward manner. This dense connectivity allows for direct information exchange between layers, enabling the network to access features from all preceding layers. This dense connection scheme promotes feature reuse and enhances gradient flow throughout the network, leading to improved performance and efficient learning.

Dataset

- The UCF-Crime dataset is a widely used dataset in the field of computer vision and video surveillance. It consists of a collection of video clips depicting various criminal activities, such as burglary, robbery, assault, and vandalism. The dataset was created by the University of Central Florida (UCF) and contains realistic scenarios with diverse lighting conditions, camera angles, and environmental factors.

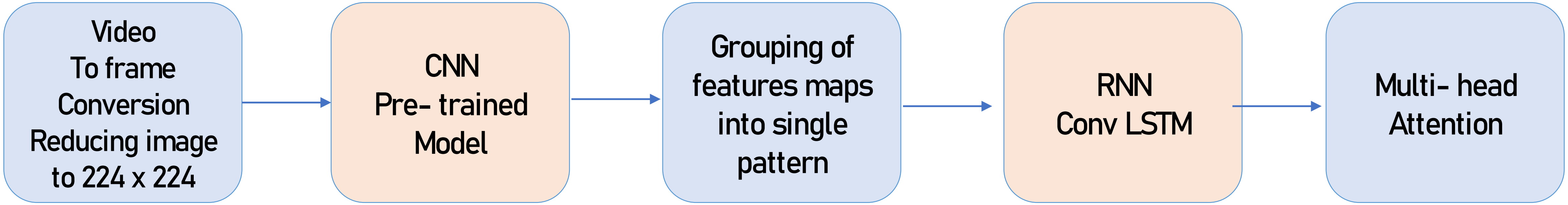


Dataset Cont’d



Binary	No. Videos	Ucfcrimes	No. Videos
Abuse	50	Abuse	50
Arrest	50	Arrest	50
Arson	50	Arson	50
Assault	50	Assault	50
Burglary	100	Burglary	50
Explosion	50	Explosion	50
Fighting	50	Fighting	50
RoadAccident	150	RoadAccident	50
Robbery	150	Robbery	50
Shooting	50	Shooting	50
Shoplifting	50	Shoplifting	50
Stealing	100	Stealing	50
Vandalism	50	Vandalism	50
Normal	950	Normal	50
Total	1900	Total	700

Proposed Model



Video
To frame
Conversion
Reducing image
to 224 x 224

CNN
Pre- trained
Model

Grouping of
features maps
into single
pattern

RNN
Conv LSTM

Multi- head
Attention

Preprocessing

- **During the initial phase, the video files are partitioned into fixed frames. This process involves determining the number of frames to be skipped when dividing a video file into n frames from its total frame count and reducing the image size into 224x224.**
- **To divide each video file into fixed frames, a specific process is followed. If a video file is to be divided into n frames out of its total number of frames, the number of frames to be skipped is calculated accordingly. For instance, if the length of a video file is 60 seconds and the video format is set to 30 frames per second, the total number of frames, denoted as m, would be 1800. Now, let's assume n is equal to 30, indicating that we aim to select 30 frames from the 1800 frames available. In this case, each frame is selected after skipping 60 frames. Once the frames are selected, the difference between each frame and its adjacent frame is calculated to capture the spatial movement for each input.**

CNN

- **Convolutional Neural Networks (CNNs) are a type of deep learning model widely used in computer vision tasks, particularly for image recognition and analysis. CNNs are inspired by the visual processing mechanism of the human brain and are designed to automatically learn hierarchical representations of visual data.**
- **CNNs have achieved remarkable performance in various computer vision tasks, including image classification, object detection, semantic segmentation, and facial recognition. They have been successfully applied in diverse fields such as autonomous driving, medical imaging, surveillance, and natural language processing with image inputs.**
- **Overall, CNNs have revolutionized computer vision by enabling machines to perceive and understand visual information, leading to significant advancements in image analysis and pattern recognition tasks.**

CNN- pre-trained model

- **A pre-trained model is a deep learning model that has been trained on a large dataset and contains learned parameters or weights that capture the knowledge from the training data. Due to difficulties in collecting and labeling anomalous events, we will use a transfer Learning in our model . Therefore, we will pre-train the model on the ImageNet dataset, including 1000 image categories . Therefore, By using a pre-trained CNN model, the model has already learned to extract meaningful features and representations from a diverse range of videos**

RNN

- Recurrent Neural Networks (RNNs) are a class of artificial neural networks designed to process sequential data, such as time series, text, or speech. Unlike feedforward neural networks, RNNs have feedback connections that allow information to persist and be carried across different time steps
- The fundamental unit of an RNN is a recurrent cell, which can maintain an internal memory state. At each time step, the cell takes an input and combines it with the previous state to produce an output and update the current state. This recurrent structure enables RNNs to capture and model dependencies over time, making them suitable for tasks that involve sequential information.
- In summary, RNNs are neural network architectures specifically designed for processing sequential data. They can model dependencies over time and have been successfully applied in various tasks that involve sequential information, ranging from natural language processing to time series analysis.

RNN- ConvLSTM

- **By using ConvLSTM, the violence detection model can effectively model the motion and dynamics present in video sequences, enabling it to recognize patterns and behaviors associated with violent activities. The recurrent nature of ConvLSTM allows the model to retain information from past frames while considering the current frame, facilitating the understanding of temporal context in violence detection**
- **The model processes the video frames in a sequential manner, capturing both spatial and temporal information. Based on the learned patterns, it predicts whether the video contains violent behavior.**

Multi– head attention

Multi – head attention

The main purpose of an attention mechanism is to assign different weights or "importance" to different parts of the input data. This allows the model to focus more on the parts that are more relevant for the prediction task, while paying less attention to less relevant parts. This can greatly improve the performance of the model, especially in tasks where only a small part of the input data is relevant for the prediction, like in our case where the accident only happens in a small part of the video.Overall, adding an attention mechanism to our ConvLSTM model can improve its ability to focus on the crucial frames in a video where an accident takes place, potentially leading to more accurate accident detection

Reference

- 1. V Singh et al. Procedia Computer Science, 2020 - Elsevier
- 2.Soheil Vosta et al. 2022 by the authors Licensee MDPI, Basel, Switzerland.
- 3.JC Wu et al. European Conference on 2022 - Springer
- 4. Chen et al. Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 1. 2023.
- 5.Cherian et al. Computational Vision and Bio-Inspired Computing: ICCVBIC 2020. Singapore: Springer Singapore, 2021. 223-230.
- 6.Sultani et al. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

Thank you