Ain Shams University
Faculty of Computer & Information Sciences
Computer Systems Department

# VIDEOS FILTRATION THROUGH DETECTING VIOLENT CONTENT IN VIDEOS

June 2021

**Ain Shams University**
**Faculty of Computer & Information Sciences**
**Computer Systems Department**

# VIDEOS FILTRATION THROUGH DETECTING VIOLENT CONTENT IN VIDEOS

**By**

Ahmed Mohamed Ibrahim [Computer Systems]

Ahmed Sherif Morgan [Computer Systems]

Ahmed Saeed Abd El-gawad [Computer Systems]

Mohamed Mahmoud Abdel Latif [Computer Systems]

Ahmed Mohamed Ahmed Mohamed [Computer Systems]

**Under Supervision of**

**[Dr. Noha Ali Abd Elsabor]**
Computer Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

**[T. A. Esraa Tarek]**
Computer Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University.

**June 2021**

# Acknowledgements

All praise and thanks to ALLAH, who provided me the ability to complete this work. we hope to accept this work from me.

we are grateful of *my parents* and *my family* who are always providing help and support throughout the whole years of study. we hope we can give that back to them.

we also offer my sincerest gratitude to my supervisors. *Dr Noha Ali Abd Elsabor and T.A Esraa Tarek* who have supported me throughout my thesis with their patience, knowledge, and experience.

Finally, we would thank my friends and all people who gave me support and

# Abstract

Violence is typically seen as a public health and social problem. Therefore, it has become more and more crucial to develop.
a technology-based application to tackle with this problem. Then, a question coming up is how to detect violent behaviors in video surveillance. In this study, we propose a deep neural network by using the pre-trained model—VGG16 combined with convolutional neural network (CNN) and long short-term memory (LSTM) deep learning models. The spatial features are extracted by pre-trained model VGG16. Training our CNN model based on spatial features will get the temporal features. Then, these features are fed to the LSTM to classify the video-based activities into specific categories (violence or non-violence behaviors). The experiments are conducted on Hockey Fight dataset, and the accuracy score is used to evaluate our model. By combination of the superior features of our own CNN and LSTM models, our proposed model has obtained encouraging experimental results much more than the others.

**Keywords** Violence Detection · VGG16 · Convolutional neural network (CNN) · Long short-term memory(LSTM)

# Table of Contents

# Chapter One:

# Introduction

# Chapter 1

# Introduction

## 1.1: **Problem Definition**

Violence is the act of using physical force against another person which either results in or likely to result in injury, psychological harm or even death. Physical violence can be the culmination of conflicts. Scientists agree that violence is inherent in humans. For prehistoric people, there has been archeological evidence proving that both violence and peace are early human characteristics. Over the past decades, public violence has been perceived to increase dramatically. In Vietnam, it is attracting more attention and concern to incidents of school violence, and it is now becoming a serious problem. According to data from the Ministry of Education and Training, in one school year, there were nearly 1600 cases of school fights inside and outside the school premises. According to some statistics, about 5200 students were involved in a fight, and 11,000 had a school dropout because of fighting. That said, it is important to design an application to detect violence behaviors automatically from surveillance cameras to prevent the serious consequences. However, almost all the current systems require manual human inspection of these videos for identifying such scenarios, which is practically infeasible and inefficient. Owing to those reasons, in this research, we propose the system which could detect violent behaviors automatically from videos.

## 1.2: **History**

Previously To rate movies, a stochastic model was proposed to capture features such as violence, profanity, etc. This model worked with a shot transition detection model to capture the amount of motion present in a scene and hence would be unable to distinguish an action movie from a game such as basketball. Researchers in also tried to capture the degree of motion present in a scene by looking at the temporal activity and length of shots along with audio cues. Their system needs manual intervention for creation of audio samples to detect sounds associated with violence. Both approaches above suffer from a fundamental problem: analyzing statistics of activity in complete scene rather than at object level, they only can ascertain that the movie genre is violent. They cannot tell us who is hitting who? Are there any objects being used? All of these questions can be answered by looking at violence from the object level. Here, we present an approach to analyzing violence at the object level. We exploit the motion trajectory information of a person during violence to calculate jerk (reaction of the hit person). We also compute the orientation of arms and legs to draw certain inferences based on their motion patterns over time.

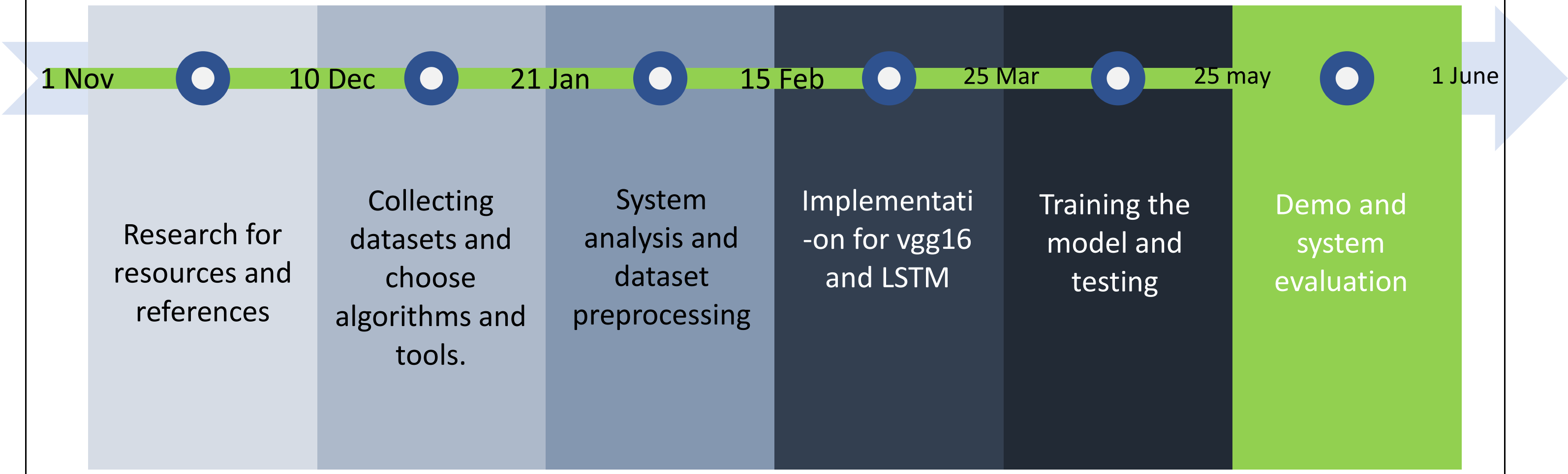
## 11.3: **Motivation**

Violence detection has been a popular topic in recent human action recognition research, particularly in video surveillance. One of the difficulties with human action recognition in general is the classification of human activity in real time, almost instantaneously after the action has taken place. This difficulty escalates when dealing with surveillance video for several factors including the quality of surveillance footage is diminished, lighting is not always guaranteed, and there is generally no contextual information that can be used to ease detection of actions and classification of violent versus non-violent. Furthermore, for violent scene detection to be useful in real world surveillance applications, detection of violence needs to be fast in order for a quick intervention and resolution. In addition to poor video quality, violence can occur in any given setting at any time of day, therefore, a solution will have to be robust to detect violence no matter the conditions. Some settings for video surveillance where violence detection can be applied includes the interior and exterior of buildings, in traffic, or on police body cameras. These use cases provide the motivation for this thesis: the ability to detect violence rapidly and accurately in real time in multiple settings.

## 1.4: **Objectives**

The goal of violence detection is to automatically and effectively determine whether the violence occurs or not within a short video sequence. In the field of video-based violence detection, it is difficult to capture effective and discriminative features as a result of the variations of human body.

## 1.5: **Time plan**



1 Nov

10 Dec

21 Jan

15 Feb

25 Mar

25 may

1 June

Research for resources and references

Collecting datasets and choose algorithms and tools.

System analysis and dataset preprocessing

Implementati-on for vgg16 and LSTM

Training the model and testing

Demo and system evaluation

### 1.5: **Documentation Outline**

this graduation project is composed of 5 main chapters.

1- Chapter one (this is where you are now and you know what it's about),

2- The Second Chapter is Background, in this chapter We will talk about a literature overview and Summary explanation for the topics that need to be covered to keep going in this work also We will discuss the survey done before this work started.

3- The third chapter is System Implementation in this chapter We will discusses input and output preparation, preprocessing and augmentation also We will
discuss the choices of the training parameters of the model finally We will show the results.

4- The fourth chapter is Conclusion and Future work in this chapter I will discusses advantages and disadvantages of Our project from our perspective,
and the future works on this project that We will do.

5- At the end there will be sections for list of abbreviations, the tools used and references for the resources and papers, that helped me out in this project.

# Chapter Two:

# Background

Chapter 2

# Background

in this chapter we will discuss and give a quick summary for the topics and skills needed to start working on this problem.

## 2.1**approaches to solve the problem**:

to solve the task of cyberbullying detection there is a classical approach using Artificial Intelligence (Deep learning and CNN) this work follow the deep learning end to end approach, hence We will introduce the concepts needed for this approach.

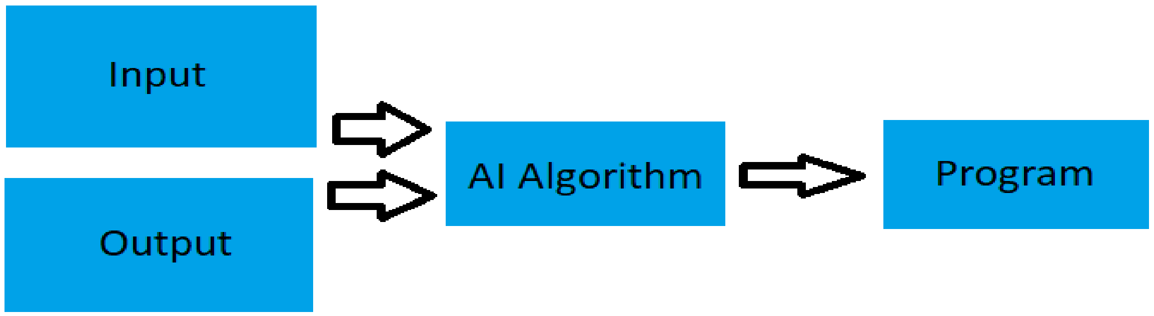
## 2.2**Artificial Intelligence (AI)**:

Artificial Intelligence is the science that enable computers to accomplish some tasks like humans, and these are kind of task that are nearly impossible to be solved using rule-based programs.

to build a normal program whatever this program is supposed to do it will be rule based, series of rules that describe an algorithm to get the output, these rules are designed based on:

- available input
- analysis of the problem

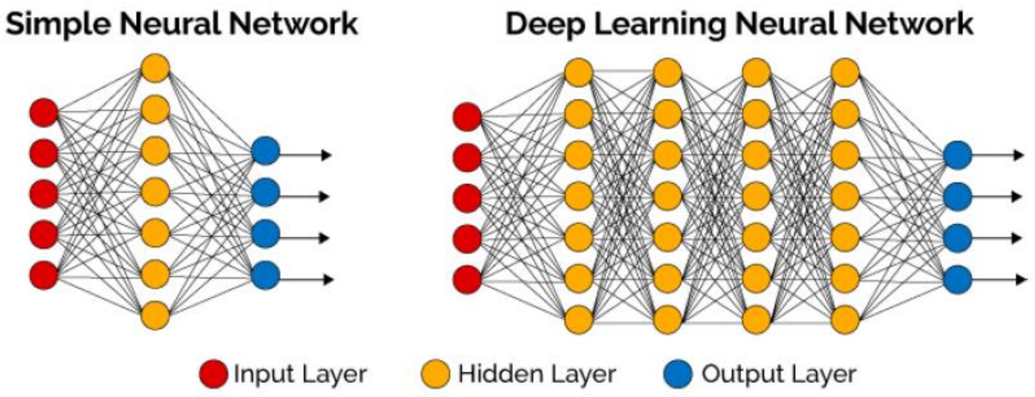
on the opposite side to develop an Artificial Intelligence model we give the model some examples of the available input and the corresponding output,



**Artificial Intelligence Model**

then using an AI Algorithm, we get the model that represent the program, then this model could take input and produce output and this process is illustrated in Above Figure.

## 2.3 **Neural Networks and Deep Learning**:



**Neural network structure**

Neural Networks are subset of machine learning that is by its turn a subset of Artificial intelligence, Neural networks try to estimate the function or the rule that will help in solving the problem in hand, the structure of Neural Networks is:

- Input Layer
- Hidden Layer
- Deep Layer

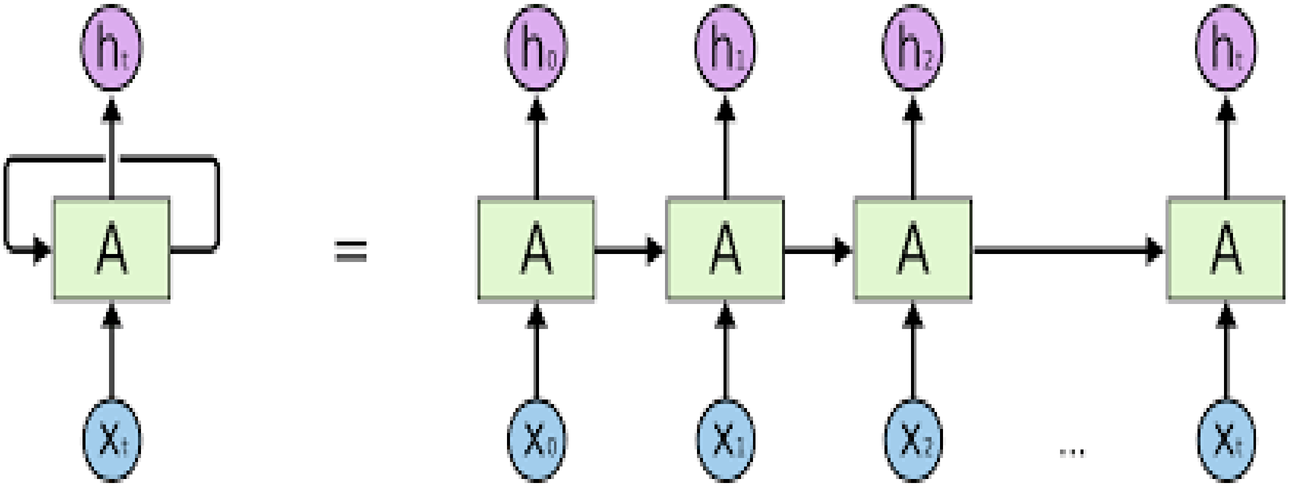
There are 2 types of neural networks:

- Shallow NN small number of hidden layers
- Deep NN big number of hidden layers

the more hidden layers the NN have the more complex functions it can represent hence, solving the problem efficiently.

## 2.4 **Recurrent neural network (RNN):**

A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence.
The term "recurrent neural network" is used indiscriminately to refer to two broad classes of networks with a similar general structure, where one is finite impulse and the other is infinite impulse. Both classes of networks exhibit temporal dynamic behavior. A finite impulse recurrent network is a directed acyclic graph that can be unrolled and replaced with a strictly feedforward neural network, while an infinite impulse recurrent network is a directed cyclic graph that cannot be unrolled.

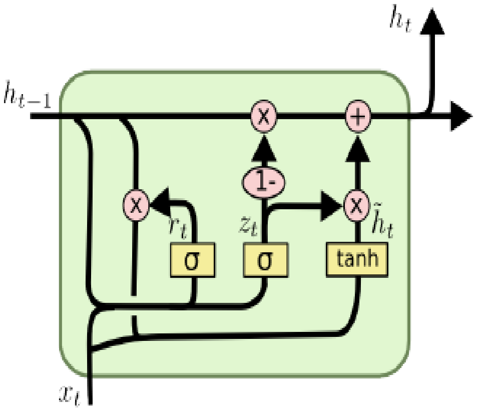


**RNN Layer**

## 2.5 **LSTM**:

Long Short-Term Memory is a special type of recurrent Neural Networks it is used to capture the long- and short-term temporal dependence between input instances.

capturing the temporal dependence is very important to the prediction task, to predict the violence we cannot rely only on the current input of this Frame only but the previous ones also, LSTM can combine theses information and produce the best possible prediction.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

**LSTM layer with the inner gates illustrated**

LSTM works throw some components called gates, they are a small Neural Networks inside the LSTM layer. they are all combined together works on deciding how much information to keep from the previous history of the input instances and how much information is used from the current input.

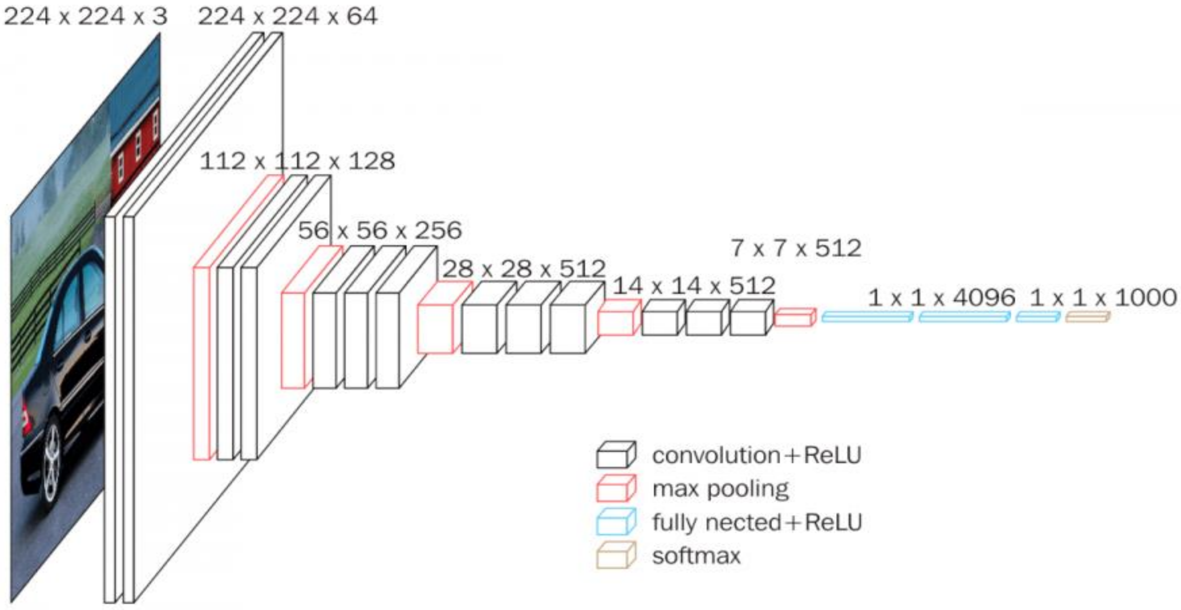
specific role of each gate could be understood from the input and the activation layer of the gate.

## 2.6 **VGG16:**

VGG16 (also called Oxford Net) is a convolutional neural network architecture named after the Visual Geometry Group from Oxford, who developed it. It was used to win the ILSVRC2014 (Large Scale Visual Recognition Challenge 2014)) competition in 2014. It still considered to be an excellent vision model.

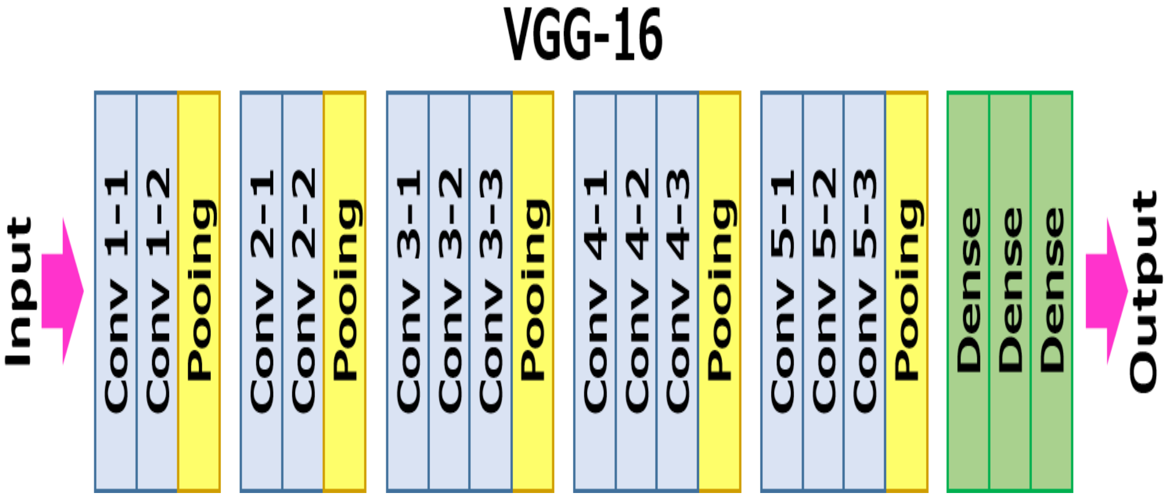
VGG-16 is a convolutional neural network that 16 layers deep. The model loads a set of weights pre-trained on ImageNet. The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes.

The default input size for VGG16 model is 224 x 224 pixels with 3 channels for RGB image. It has convolution layers of 3x3 filter with a stride 1 and maxpool layer of 2x2 filter of stride 2.



## The Architecture of VGG16

- The Kernel size is 3x3 and the pool size is 2x2 for all the layers.
- The input to the VGG16 model is 224x224x3 pixels images.
- two convolution layers with each 224x224x64 size.
- A pooling layer which reduces the height and width of the image to 112x112x64.
- Then we have two conv128 layers with each 112x112x128 size after that we have a pooling layer which again reduces the height and width of the image to 56x56x128.
- three conv256 layers with each 56x56x256 size, after that again a pooling layer reduces the image size to 28x28x256.
- three conv512 layers with each 28x28x512 size, after that again a pooling layer reduces the image size to 14x14x512.
- three conv512 layers with each 14x14x521 layers, after that, we have a pooling layer with 7x7x521.
- two dense or fully connected layers with each of 4090 nodes. and at last.
- Then we have a final dense or output layer with 1000 nodes of the size which classify between 1000 classes of image net.



## 2.7: **VGG16 Dataset (ImageNet)**:

ImageNet is an image database with a total of 14 million images and 22 thousand visual categories. As it is publicly available for research and educational use, it has been widely used in the research of object recognition algorithms and has played an important role in the deep learning revolution, ImageNet has been mostly used for researching object recognition algorithms on the subset of 1000 categories.

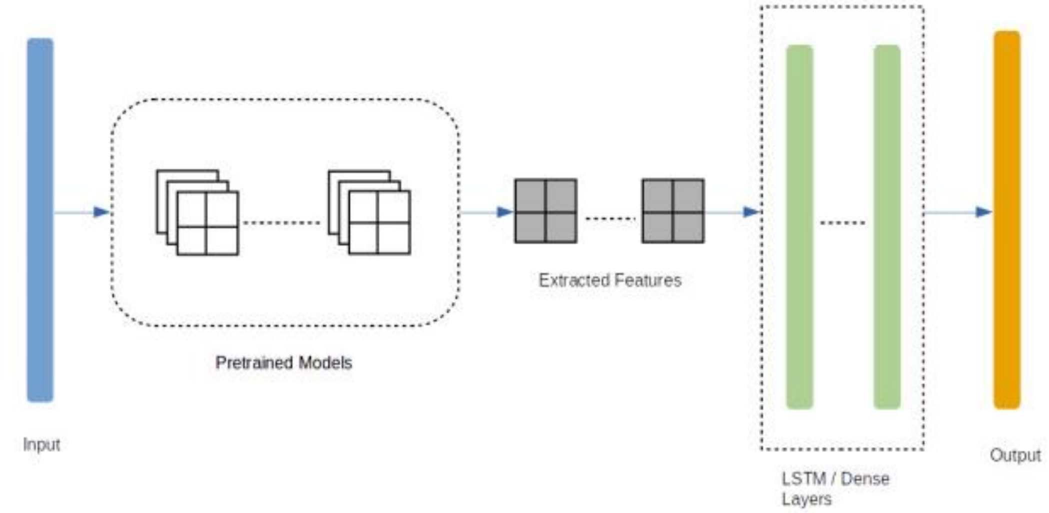
# Chapter 3:

# System Architecture & Implementation

3.1 **System Architecture**
3.2: **Dataset**
3.3 **Implementation**
3.4: **results**

## 3.1 **System Architecture**:

The following figure shows the main architecture of our violence Detection system, After the preprocessing step we used the output as an input for the pretrained model in our case its VGG16,
Then we use the output features as an input for the LSTM model.



## 13.2: **Dataset**

We used the Hockey Fight dataset  as our experiments. This dataset recorded Hockey players’ matches with 500 violent and 500 non-violent clips. We divided the dataset into two parts: Train, Test with ratio is Train: 800 clips (400 violent and 400 non-violent clips), Test: 200 clips (100 violent and 100 non-violent clips).

## 3.3: **Implementation**

### 3.3.1**: preprocessing**

In this section I will discuss the input and output preparation for violence detection system

1. Extract 20 frame for each video
2. Resize each frame to a size of 224 × 224.
3. Normalize each frame to be in the range of 0 to 1.
4. For each video we make a label (labeling)

### 3.3.2: **VGG16 (feature extraction)**

we have extracted features from the frames of the videos using VGG16. The extracted features are being fed into a fully connected layer. First, we input and process 20 video frames in batch with the VGG16 model. Just prior to the final classification layer of the VGG16 model.
After that we save the Transfer Values to a cache-file.
The reason for using a cache-file is that it takes a long time to process an image with the VGG16 model. If each image is processed more than once, then we can save a lot of time by caching the transfer-values.

When all the videos have been processed through the VGG16 model and the resulting transfer-values saved to a cache file, then we can use those transfer-values as the input to LSTM neural network.

Then we will train the second neural network using the classes from the violence dataset (Violence, No-Violence), so the network learns how to classify images based on the transfer-values from the VGG16 model.

### 3.3.3: **LSTM** (classification)

When defining the LSTM architecture, we have to take into account the dimensions of the transfer values. From each frame the VGG16 network obtains as output a vector of 4096 transfer values. From each video we are processing 20 frames so we will have 20 x 4096 values per video. The classification must be done considering the 20 frames of the video. If any of them detects violence, the video will be classified as violent.

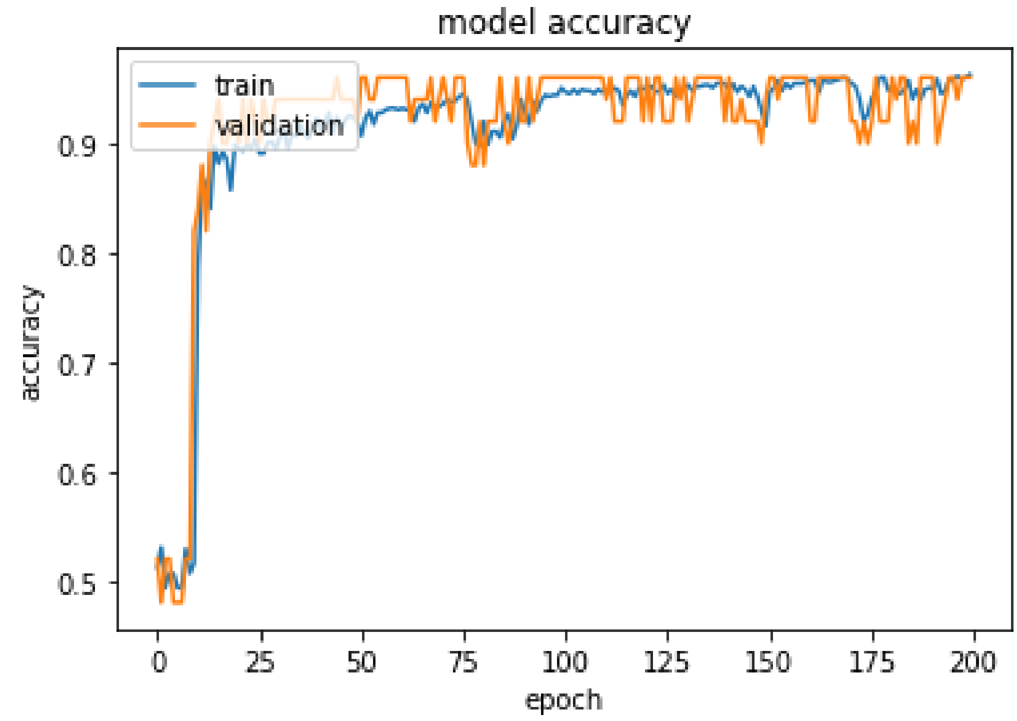
The first input dimension of LSTM neurons is the temporal dimension, in our case it is 20. The second is the size of the features vector (transfer values).

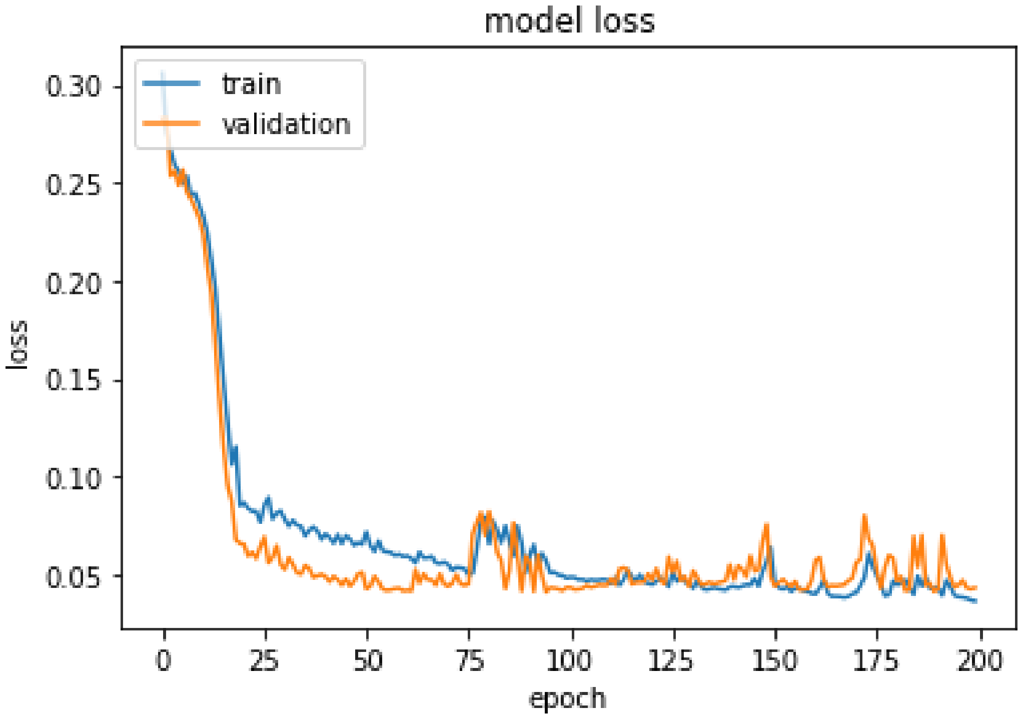
## 3.4: **Results**

Our model based on VGG16, and LSTM was first trained on one of the benchmark datasets, namely Hockey .the Hockey’s dataset consists of 1000 videos. The datasets have an equal number of Violence and Non-Violence videos, The code has been written in Python.

We have used Keres deep learning library with TensorFlow backend to implement our desired models. We have trained all the models with a common training set and tested them all. We let the model run for 200 epochs and reported their accuracy metrics for training and testing.

| Model | Accuracy |
|---|---|
| VGG16+LSTM | 94% |

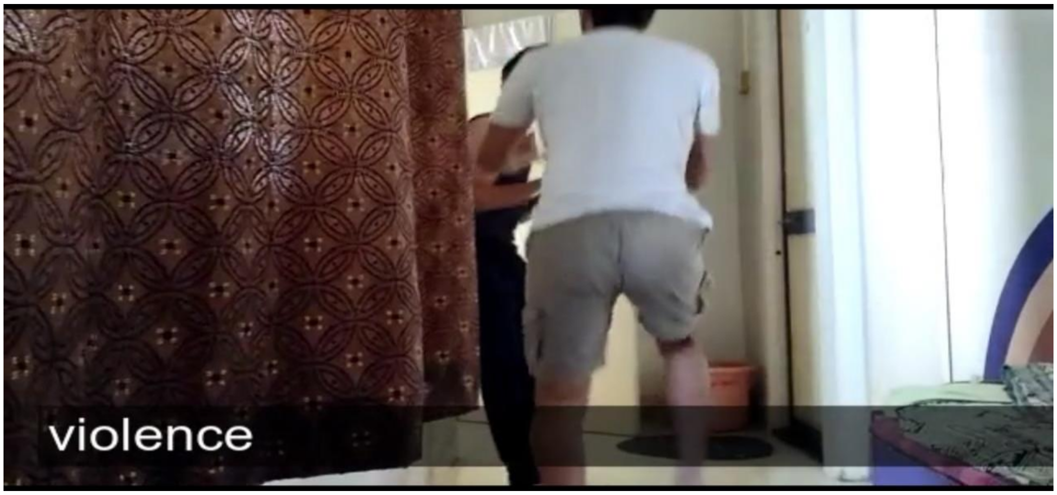

model accuracy
train
validation
accuracy
0.9
0.8
0.7
0.6
0.5
0
25
50
75
100
125
150
175
200
epoch

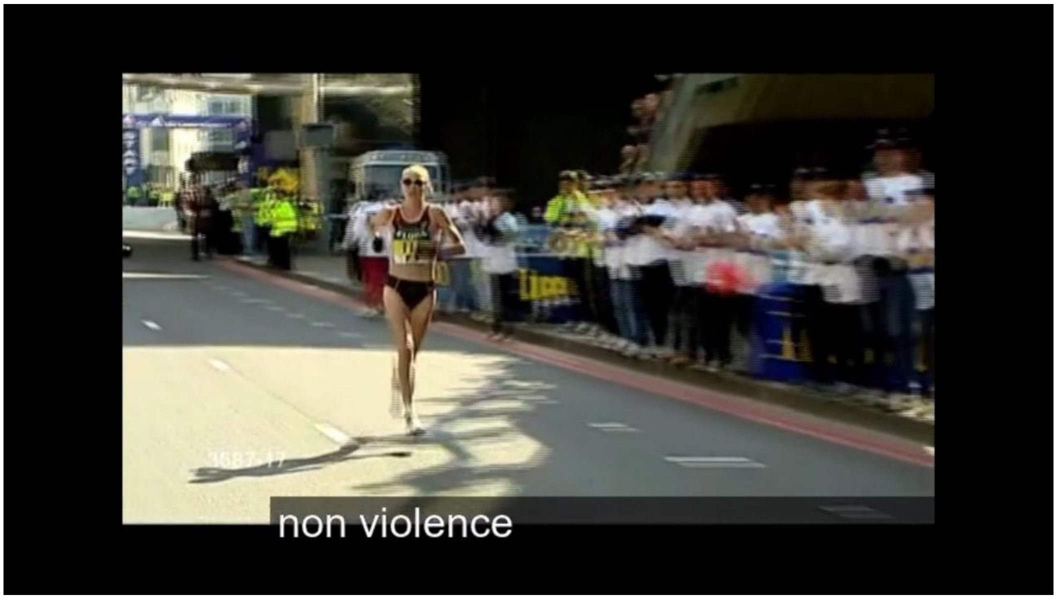

model loss
train
validation
loss
0.30
0.25
0.20
0.15
0.10
0.05
0
25
50
75
100
125
150
175
200
epoch

## 3.5: Testing Process

After finishing the training process of the model we tested it using the 20% of the dataset (never seen by the model), we obtained an accuracy of 94%.

*Test samples:*





# Chapter 4:

# Conclusion and Future work

4.1: **Conclusion**
4.2: **Future Work**

## 4.1: **Conclusion**

The results from our experiment demonstrate the effective use of CNN+LSTM architecture, for training the model over two popular Hockey dataset. And it highlights that a convolutional neural network which leverages transfer learning with long short-term memory networks outperforms all the other variance of convolutional neural networks. By combining CNN with LSTM, the accuracy increases to a certain margin as compared to pure transfer learning models. The system provides a simple graphical user interface to interact with deep learning model. This study explores and dives deep into leveraging the potential of extracting salient features from the frames which then have been used in detecting violence in the videos. We have experimented with pretrained VGG16. The extracted features from each of the frames have been fed into a fully connected network. Moreover, in another experiment, extracted features from 30 frames at a time have been given to an LSTM network as an input sequence. Furthermore, a spatial transformer network has been designed to apply transformation and attention to the extracted features.

## 4.2: **Future Work**

At this point we cannot use this model on live camera just with videos that have been already saved so it will be a productive and useful to find a way to make this system runs in live camera

# Chapter 5:

# Appendix

# Tools

## 5.1: **Software tools**

- Python: is the main programming language for this project.
- keras: is an open-source deep learning library for building, training and testing the model.
- open cv: for deferent preprocessing and augmentation techniques
- matplotlib: for result visualization and data set analysis
- NumPy
- pandas: for data set loading and manipulations
- TensorFlow
- OpenCV
- moviepy

## 5.2: List of abbreviations

- **AI: Arterial Intelligence**
- **NN: Neural Network**
- **CNN: Convolutional Neural Network**
- **RNN: Recurrent Neural Network**
- **LSTM: Long Short-Term Memory**

# References

1. Violent Video Detection by Pre-trained
   Model and CNN-LSTM Approach
   Bui Thanh Hung, Vijay Bhaskar Semwal, Neha Gaud,
   and Vishwanth Bijalwan

2. Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM
   Zahidul Islam, Mohammad Rukonuzzaman, Raiyan Ahmed, Md. Hasanul Kabir, Moshiur Farazi

3. J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient violence detection using 3d convolutional neural networks," in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019, pp. 1–8.

4. Keres https://keras.io/
5. S. Hochreiter, J. Schmidhuber, Long short term memory. Neural Comput. 9(8), 1735–1780
   233 (1997). https://doi.org/10.1162/neco.1997.9.8.1735

6. Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," Image and vision computing, vol. 48, pp. 37–41, 2016.

7. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight recognition in video using hough forests and 2d convolutional neural network," IEEE Transactions on Image Processing, vol. 27, no. 10, pp. 4787–4797, 2018.

8. P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violent interaction detection in video based on deep learning," Journal of Physics: Conference Series, vol. 844, p. 012044, 06 2017.

9. P. Bilinski and F. Bremond, "Human violence recognition and detection in surveillance videos," in 2016 13th IEEE International Conference on Video and Signal Based Surveillance (AVSS). IEEE, 2016, pp. 30–36.

# List of Figures