Automate Violence Detection

Team 4: AI Mania
University of Windsor, Windsor, Ontario

Abstract

As new era of technology has been initialized in current education system which includes schools, colleges and universities, cases of physical violence has also increased. Solution to that question is automatic detection of such violence with help of pre-installed surveillance cameras. Automate Violence Detection system helps to arise alert by detecting violence from gestures and postures of one person with respect to other person with help of multiple data sets and machine learning algorithm. Through some sources and basic knowledge, project was developed to understand actual working of this system where new challenges were faced and solution to many problems were identified. We have attempted to solve those issues by using Neural Network, especially, Recurrent Neural Network. As mentioned in the book *the quest for Artificial Intelligence* by Nilsson, Neural Networks has been around since early 19th Century and began when Warren McCulloch and Walter Pitts essentially claimed a neuron as a “logical unit” (Nilsson, 2009). RNN is class of neural network for deep learning which has feedback connection unlike the traditionally used feed forward neural networks. The proposed RNN network uses VGG16 and Bidirectional LSTM (Long Short -Term Memory) to extract features from individual frame of the video, the sequence of frame features is then taken into bidirectional LSTM recurrent networks for classifier. Thus, this research can be good source to one who is studying in field related to automatic prediction system through surveillance cameras (Li, et al., 2019).

Keywords: VGG16, Bidirectional LSTM, RNN network, Video frames, Violence Detection, MVC architecture, Spring framework.

Introduction

Common problem faced in current day to day life with people is bullying and physical violence at schools/ university for the sake of jealousy, showing power or to maintain superiority. For safety and security purpose, cameras are installed at various locations like premises, rooms, or outdoor areas of respected places. This can help to monitor activities in those places covered by cameras.

As the number of CCTV surveillance have increased, it requires more people to handle and monitor each place. Ultimately, it took more time detect such activity in monitor room for 24*7 and still many incidents are missed which leads to occurrence of crime. Because of this, detecting violent activities automatically with help video surveillance system have received great attention.

As a result, need of automation system in this field elevated which can manage behaviour and actions of one person with respect to other person by monitoring, analysing, detecting level of intensity, and finally generating a security alert. This is possible with the help of machine learning algorithm, some datasets, database management system and high working hardware devices which can handle load and faster processing of video datasets. After basic needs, video frames are processed and converted into fixed fps (frames per second) size and compared with predefined datasets.

There have been excessive research going on in field of automatically analysing human behaviour. But exact definition of violence differs in multiple way based upon situation. For an instance, it can be just fist fight or there can be weapon. Even, it can mislead to scenario due to wrong prediction. For an instance, it can be a person helping another person who recently fall due to slippery surface or else it can be two friends having fun. Thus, applying limitations and updating feature one by one can perform detailing in this.

Literature Review

VGG16 and Bidirectional LSTM:

Setting algorithms with of python language is much easier. Functions are easily available and machine learning is much more compatible with python language. VGG16 is a convolution neural network which is used for large scale of datasets with many images. It was developed by K. Simonyan and A. Zisserman. This model has very high accuracy i.e., 92.7%. RNN is recurrent neural network, which was designed to work with sequence prediction problem, (Brownlee, 2018). Use of LSTM (Long Short – Term Memory) have made the task easy to train recurrent network (Naik & Gopalakrishna, 2017).

After that, combination of these two algorithms makes it simple to understand and easier for the application. As videos are monitored from different cameras, their sizes differ. So, first step is to convert every incoming video frame into equal fps (frames per second) before loaded on screen. Thus, there is requirement of high performing GPU. The other way to understand this can be through static manner. Taking some saved videos into consideration, converting it into equal fps size and same extension, it can be imported to violence detection system and checked for the violence.

Then, LSTM algorithm will help to extract features from those frames and further compared with predefined datasets. For an instance, videos converted to 10 frames per second. Thus, 1st set contains 10 frames, 2nd contains 10 and so on. LSTM will process received data concurrently. Bidirectional LSTM is used for the accuracy purpose. It is easy to understand with an example. If there are 10 frames in first set and count for accuracy is set for 4, then with help of predefined datasets (which includes multiple frames of gestures and postures of violence like – fist fight, leg kicks, pushing other person aggressively), received frames will be compared to those datasets in manner – (K,1,2,3), (1,2,3,4), (2,3,4,5), …,(7,8,9,10), (8,9,10,K), (9,10,K,K), (10,K,K,K). K is constant of just blank frame so that each frame is checked and compared to violent datasets multiple time for accuracy. Thus, detailing is done with accuracy for safety purpose.

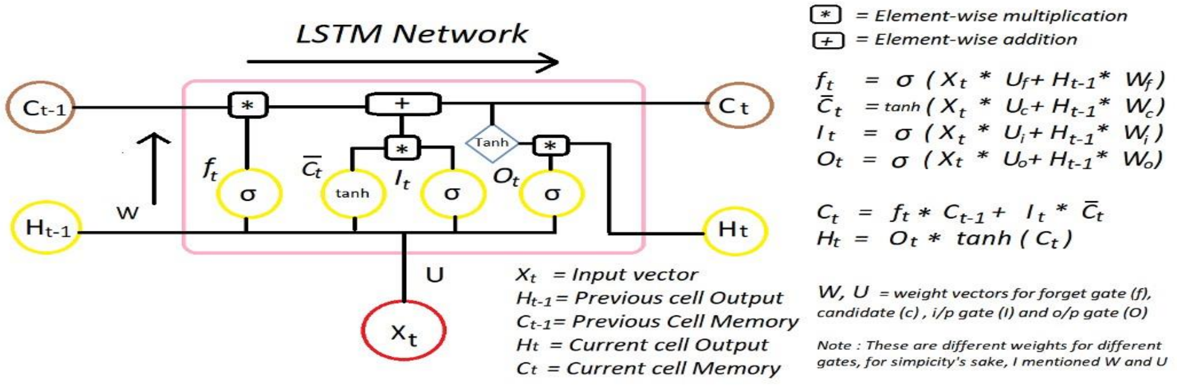
LSTM Network
C_{t-1}
C_t
H_{t-1}
H_t
W
U
X_t
f_t
$\bar{C}_t$
I_t
O_t
σ
tanh
Tanh
*
+
X_t = Input vector
H_{t-1} = Previous cell Output
C_{t-1} = Previous Cell Memory
H_t = Current cell Output
C_t = Current cell Memory
* = Element-wise multiplication
+ = Element-wise addition
$f_t = \sigma (X_t * U_f + H_{t-1} * W_f)$
$\bar{C}_t = tanh (X_t * U_c + H_{t-1} * W_c)$
$I_t = \sigma (X_t * U_i + H_{t-1} * W_i)$
$O_t = \sigma (X_t * U_o + H_{t-1} * W_o)$
$C_t = f_t * C_{t-1} + I_t * \bar{C}_t$
$H_t = O_t * tanh (C_t)$
W, U = weight vectors for forget gate (f), candidate (c) , i/p gate (I) and o/p gate (O)
Note : These are different weights for different gates, for simpicity's sake, I mentioned W and U

Figure 1 Working of LSTM Network, (Sanjeevi, 2018)

As shown in Figure 1, it can be understood easily. Two functions are used which are sigmoid function and tan function. f, I and O are single layered neural network with sigmoid activation function and C with tan activation function. C is memory state. For an instance, "frame 1 taken for comparison. Dataset 1 from predefined dataset taken for comparison. Thus, frame 1 compared with?" To predict the answer correctly, algorithm stored "Dataset 1" in memory C.

LSTM then takes previous cell memory Ct -1 and multiply it with forget gate f i.e., Ct = Ct-1 * f, (Sanjeevi, 2018)

Thus, if value of f is 0, previous memory state is completely forgotten. And if f = 1, it is completely passed to memory. It is easy to understand it with example. Suppose count value for frames to be compared is set as 4. So, for set (K,1,2,3) is checked first. Hence, there will not be any previous state. Each frame is checked once. Thus, frame K will not be passed to memory state as f = 0. For other frames, previous memory state will be K, so they will be passed to memory state. After increasing count, with current memory state Ct, new memory state from input state and C layer will be calculated. So, it will be Ct = Ct + (It * C't), (Sanjeevi, 2018)

Thus, new input will be frame number 4 for next set (1,2,3,4), and again those frames in current memory state will be compared with predefined dataset and same situation will arise for frame number 1. Thus, algorithm flows in both the direction. Previous memory state to next memory state and again same situation. Thus, Ct which is current memory state at time step t will be passed to next time step.

Thus, output of this calculation will be based on current memory state Ct but more filtered. For that, tanh function is applied to Ct and element wise multiplication is done with output gate O.

In simple, frames will be filtered, and output will be generated i.e., if violence detected or not. Final calculation will be current hidden state Ht i.e., Ht = tanh (Ct), (Sanjeevi, 2018)

In 2018, Chen in his GitHub repository – Keras Video Classifier, explained working modules of both algorithms. With combination of these two algorithms i.e., VGG16 and LSTM bidirectional, accuracy vs loss graph was measured and following results were obtained.

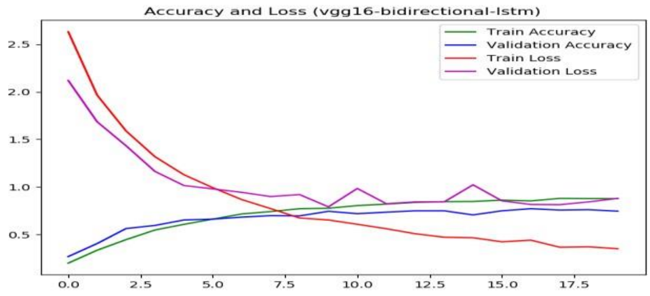

Figure 2 Trained Graph of VGG 16 with Bidirectional LSTM

- The accuracy obtained for VGG16 and bidirectional LSTM was around 89% for training and 77% for validation, (Chen, 2017)

Working of algorithm and system

Working of Algorithm

The profound expectation neural system is based on the Convolutional Neural Network (CNN) and a variation of the Recurrent Neural Network (RNN)- Long Short-Term Memory (LSTM). In the beginning, data from videos captured can be get through CNN and to compare it with predefined datasets, videos of predefined datasets can be obtained through LSTM algorithm.

At last, the detection can be done whether frame contains violence or not. These experimental results help to check accuracy of the system and comparing it with datasets proper results can be achieved (Nilsson, 2009).

Working of system

Working of system can be understand by dividing it into different phases.

The hardware architecture of physical violence is shown below in the figure. Hardware used for static experiment are cameras, accelerometer, gyroscope, RAM, CPU, and email module. Process begins by capturing videos through surveillance cameras. So, it is necessary to reduce large sized videos to small sized video frames for gesture detection. Videos are generally saved in RAM and later into secondary storage. Videos stored in RAM can be reduce to specific fps size through any open-source application like Filmora. CPU is used for processing of algorithm. As LSTM algorithm is trained bidirectionally, each video frame is compared multiple times with predefined dataset for accuracy and it is already explained above. If any frame matches with predefined dataset then flag is turned to positive and with help of email module, those frames can be sent as an alert email to be registered email address.

Even in single institution, there are multiple cameras, and each camera gathers huge of amount of data. Thus, to process that large data, there is need of high performing GPU which can handle management and processing of such large data. Parallelly, to understand working of system without GPU, it can be done in static manner.

Static Processing Without GPU

For experimental purpose, first create some datasets of violence and non-violence videos. Dataset1 have 10 videos of violence with 25 fps. In those 10 videos, include different types of physical violence like kicking, fist fight, pushing aggressively, etc. Dataset 2 have 5 videos of non-violence with 25 fps. In these 5 videos, include videos of friends moving together, chilling and doing fun, two people studying together, etc. After creating datasets, create experimental videos of fight.

Comparison of captured video having violence with predefines datasets.

For an experiment, create video of 15 seconds of two people fighting with each other where gestures include fist fight and kicking. Reason for video of 15 seconds is as GPU is not included with system, so without GPU, short video will be processed speedily compared to long videos. Moreover, it is easy to convert those videos to respective fps size and specific format. Thus, all videos will be of specific size and format and easily compared with datasets.

If violence is detected in frame number 4 of 1st 10 frames by comparing it with predefined dataset then, using LSTM bidirectional, it is checked again as explained inVGG16 and RNN for accuracy purpose. Thus, it does not make false alert. Hence, alert email can be sent to registered email address

Comparison of captured video having non - violence with predefines datasets

For an experimental purpose, create normal video of 15 seconds of two friends meeting, shaking hands and walking together. As it is short video, it can be processed faster.

After that, convert it into specific format and respected fps. Datasets contains gestures and postures with respect to violence.

Thus, after processing of that video and comparing it to predefined datasets, there will not be any frame which turns flag into positive and video will be compared till last frame. Thus, no violence will be detected, and no alert email will be sent.

Comparison of multiple videos with predefined datasets

Further to obtain more accuracy, create video of 30 seconds where after few seconds of beginning, two people starts fighting. After fight of 10 seconds, third person arrive there to stop the fight. At the end, both people stopped fighting and did hand shaking.

Again, perform same process of fetching video in system and converting it into respected size and format. After processing it and comparing with predefined datasets, response getting back to MVC architecture will turn the flag to positive for respected frames which will contain violence. For an instance, from frame number 2 of 11th set of 4 frames (set count of 4 fps), violence began. Thus, from that frame till frames until 10 more seconds, violence will be detected for whichever frame matches violence gesture and posture with predefined gestures and postures. Alert email will be sent to registered email address. From 21 seconds till end of video, no violence will be detected. Hence, at that time flag will not turned to positive.

Email Alert Module:

Email module can be defined with the help of MVC architecture in Java language and violence detection function can be called in this MVC architecture. Creating and managing services with help of Spring MVC is quite delicate. Naming convention should be followed strictly so that it does not create confusion while calling services in project. Creating same device as server on which experiment is done, it will slower down overall process. Thus, on one hand server services works and on other hand, MVC architecture works and algorithm set with help of python language is called to detect for violence. Hence, if violence is detected, flag set in MVC architecture, turns to positive and email alert function is called. After that, email module can be set with two email addresses i.e., 1st of system designed to send an alert email and second to the registered email, to whom alert can be send. Finally, email alert can be sent to registered email address.

If violence is not detected, no alert email will be sent as flag will not turn to positive and system will proceed checking next consecutive video frames.

Experimental Setup and Methodology

Implementation Environment

Python and java languages will be used for the project. Eclipse and PyCharm IDE are the integrated development environment (IDE) used. Eclipse uses the spring framework and

PyCharm is used for machine learning algorithm. Flask and Spring for Web development.

Coding Standard for Spring Framework Structure

Coding standards are rules and regulations for coding. It simplifies developed code and help to understand the code easily. It includes multiple things like how to name variables, how code should be laid out, where to put comments and work of functions to be carried out. This section describes the coding standards used in program. For multiple function belonging to same class, coding standard is followed as "functionName_number" i.e., first word of function name should contain all letters in lowercase, continuing second word, which should have first letter in uppercase and rest all letters in lowercase. Underscore followed by a number should be used for functions with same name. For an instance: loginAdmin_1, loginAdmin_2.

Further coding standards used are as follow:

Naming Conventions

The name of variable that used in script represents the content or purpose or role of the Variable and are defined with the length of seven to eight characters. Variable names consist of a data is used in it. If it is a string, then the prefix of the variable is 'string' else if integer then 'int'.

Comments

The comments should describe what is happening, how it is being done, what parameters mean, which global are used and which are modified, and any restrictions or bugs. Standards adopted for comments are Every script should begin with a comment block, which describes the scripts purpose; any Arguments used (if applicable), and return values (if applicable), inputs-outputs, and name of Script. Comments can also be used in the body of the script to explain individual sections or lines of code and for variable definition or declaration. Each part of the project has a specific comment layout. e.g., Line comments (//…), block comments (/*….*/) etc.

Stored Procedure

Uses for stored procedures include data-validation (integrated into the database) or access control mechanisms. Furthermore, stored procedures can consolidate and centralize logic that was originally implemented in applications. Used RDBMS for storing the data into the data tables. Front-end validation is provided using the java script. User data is stored into relational database using Hibernate Query Language. The web application uses MVC architecture. It includes Model, View and Controller. The data entities are provided into VO files. A model contains the data of the application. A data can be a single object or a collection of objects. A controller contains the business logic of an application. Generally, the @Controller annotation is used to mark the class as the controller. A view represents the provided information in a format. Generally, JSP+JSTL is used to create a view page, (Dudnik, 2014).

Drawbacks of VGG16

Drawbacks of VGG16 is to train it. It is very slow to train. Moreover, the network architecture itself weighs more with respect to disk/bandwidth. Due to its depth and fully connected network, VGG16 almost weighs more than 530 mbs size. (Hassan, 2018)

Alternate Method

Alternative method can be through use of other hardware and fuzzy algorithm. Fuzzy algorithm can be used for detecting bullying activities which follows some basic concepts. As a view, any activity take place in horizontal manner, its velocity and speed occur in horizontal manner. Thus, gyroscope vector is also calculated in horizontal manner. For size and length, vertical vector is measured. (Ye, Ferdinando, Seppanen, & Alasaarela, 2014).

$$\overrightarrow{\mathrm{Acc}_{\mathrm{Hori}}} = \overrightarrow{\mathrm{Acc}_{x}} + \overrightarrow{\mathrm{Acc}_{z}},$$

$$\overrightarrow{\mathrm{Gyro}_{\mathrm{Hori}}} = \overrightarrow{\mathrm{Gyro}_{x}} + \overrightarrow{\mathrm{Gyro}_{z}}.$$

Both algorithms together represent combined horizontal acceleration and gyro vectors. Y-axis represents vertical vector.

Then, frequently changeable videos are taken into consideration where signals are measured through graph. If fall is identified, it is bullying, or else algorithm again hit detection module to check for any activity of violence. With a threshold set, it confirms that it is bullying and not any false alert. Thus, the flow of system here starts from monitoring videos through cameras. After that, for every incoming video, accelerometer and gyroscope will perform their works and data will be provided to the RAM for processing. Processed videos will be passed to CPU for processing algorithm and comparing those extracted videos with predefined datasets. If violence is detected, flag will be turned to positive and alert email will be generated. Flag will remain positive till each frame detected with violence and once processed video frames do not match as violence with predefined dataset, flag will turn back to negative for checking of further data.

Conclusion

This application can detect one or more individuals engaged in violent activities through cameras. Action recognition techniques that have focused largely on individual actors and simple events can be extended to this specific application. This application be mainly useful in video surveillance scenarios like in prisons, psychiatric or public places.

References

Brownlee, J. (2018, 7 23). *When to Use MLP, CNN, and RNN Neural Networks*. Retrieved 12 2, 2020, from Machine Learning Mastery: https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/

Chen, X. (2017, 12 12). *keras-video-classifier*. Retrieved 12 2020, from GitHub: https://github.com/chen0040/keras-video-classifier

Dudnik, O. (2014). *Spring MVC.* Retrieved 12 2020, from Academia: https://www.academia.edu/42929889/Spring_MVC

Hassan, M. U. (2018, 11 20). *VGG16 – Convolutional Network for Classification and Detection*. Retrieved 12 2020, from Neurohive: https://neurohive.io/en/popular-networks/vgg16/

Li, A., Miao, Z., Cen, Y., Mladenovic, V., Liang, L., & Zheng, X. (2019). Global Anomaly Detection Based on a Deep Prediction Neural Network.

Naik, A. J., & Gopalakrishna, M. T. (2017). Violence Detection in Surveillance Video-A survey. *International Journal of Latest Research in Engineering and Technology (IJLRET)*. Retrieved 12 2020

Nilsson, N. (2009). *The quest for artificial intelligence : A history of ideas and achievements.* Cambrige University Press. Retrieved 11 2020, 25, from https://ai.stanford.edu/~nilsson/QAI/qai.pdf

Sanjeevi, M. (2018, 1 12). *Chapter 10.1: DeepNLP — LSTM (Long Short Term Memory) Networks with Math.* Retrieved from Mediium: https://medium.com/deep-math-machine-learning-ai/chapter-10-1-deepnlp-lstm-long-short-term-memory-networks-with-math-21477f8e4235

Ye, L., Ferdinando, H., Seppanen, T., & Alasaarela, E. (2014). Physical Violence Detection for Preventing School Bullying. *Advanced in Artificial Intelligence*. Retrieved 12 2020

Appendix A

Workload distribution

Team Effort:

1. Study of domain.
2. Problem Identification.

Anim Malvat:

1. Database management for user details
2. Implementing email module for sending alerts to users.

Ashmit Samrha:

1. Front-end work: HTML5, CSS3, JavaScript
2. Preparing Presentation

Nikhil Kothari:

1. Training the ML model using python.
2. Helping with Research Report.
3. Helping with preparing presentation.

Rushabh Rathod:

1. Integrating the trained ML Model with the web app using Python.
2. Research Report writing.

Yash Joshi:

1. Data collection and creating datasets.
2. Spring Maven Project.

Appendix B

Recurrent Neural Networks and Long Short-Term Memory

We have used Recurrent Neural Network (RNN) that is just an update of the Neural Network. A recurrent neural network is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behaviour. Derived from feedforward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs. Also, as RNN remembers the prior leaning and make use of those in making unlike the normal neural network which can implement the same using multiple simple neural networks, but the essence of a sequence, that is, the dependence and influence of sequence values are lost if individual components of sequence while using multiple simple neural networks.

Traditional neural network deals with vanishing gradient problem, where the feed forward neural network cannot send the needed gradient information during back propagation. To overcome this problem of RNN, Long Short-Term Memory (LSTM) neural network is used which can process entire sequence of data. LSTM allows gradients to flow unchanged or without vanishing or exploding to an extent. There are several successes achieved using RNN with LSTM units.

Convolutional Neural Network and VGG16

A Convolutional Neural Network (CNN) falls in the Deep learning category which is used for image classification. CNN can capture Spatial and Temporal dependencies in an image by using layers of filter in the neural network and that is how it is different from the RNN. There are hidden layers in the CNN that convolves with a multiplication or other dot product. The dot product of each calculation is used as input for next layer.

It is a Convolution Neural Network architecture for computer vision and image processing. The 16 in VGG16 means that the model has 16 layers that have weights.

Appendix C

Spring framework

Spring is an open-source Java platform framework for developing Java web application on top of the Java EE platform. It enables the use of Plain Old Java Object in the enterprise-class application. It provides several ORM frameworks, logging frameworks and other technologies that helps make the development quick and easy. Furthermore, it provides MVC based development framework and consistent transaction management interface which can easily scaled up.

Flask Framework

Flask is also a web application framework which is written in python. It is more of a micro web framework as it does not require tools or libraries. It does not provides inherit support for database abstraction or form validation or any other component, but it supports extensions that can be added to implement the features for object-relational mapper, form validation, upload handling, and various other common tools.