VIOLENCE DETECTION IN SURVEILLANCE VIDEOS USING 3D-CNN

Syed Muhammad Raza Naqvi

Habib University,
Karachi, Pakistan.
sn03805@st.habib.edu.pk

Muhammad Ahmed

Habib University,
Karachi, Pakistan.
ma04007@st.habib.edu.pk

ABSTRACT

Automated Violence Detection in surveillance videos is both a challenge and crucial as a technology. A lot of techniques have been developed to detect violence in videos based on machine learning, feature descriptors and deep learning. Still this field remains an active area of research whether it is in form of dataset generation, developing new techniques or hybridizing existing methods. In this project, we have attempted to approach problem of violence detection in videos using deep learning. 3D-CNN architecture has been used to classify videos while most recent dataset RWF-2000 has been used to train the model.

Index Terms— Surveillance, Violence, 3D-CNN, RWF-2000.

1. INTRODUCTION

Today, with the increasing number of surveillance cameras being installed in public places, detection of violence arises as a challenge. It is obvious that a human can not efficiently monitor a large number of cameras and chances of missing important events such violence are high. That is why automated violence detection has emerged as an active area of research.

Initially techniques were based on extracting features according to various descriptors and using them to classify videos as violent or non-violent. Then with advances in deep learning technologies, CNN became popular for classification of videos although availability of large public dataset remains a challenge. Now, hybrid techniques are emerging which combine feature extraction using descriptors and deep learning methods.

In this project, we have attempted to approach this problem using deep learning techniques. 3D-CNN approach has been used, rationale of which will be discussed later. Most recent and large dataset of surveillance videos RWF-2000 has been used to train 3D-CNN model.

The rest of the paper is organized as follows. Section 2 is a short literature review in this regard. Section 3 describes the opted methodology. Section 4 describes results obtained after experimentation. Finally, Section 5 concludes the paper by detailing possible future directions to proceed in this area.

2. LITERATURE REVIEW

We began our literature review by reviewing existing datasets for violence detection. Ming et al. [1] reviews existing datasets and introduces a new dataset as well. Existing datasets are Movie fight, containing 200 videos from movies, Crowd Violence, containing 246 videos of crowded places, Hockey fight, containing 1000 videos of Hockey fight and non-fight scenes. Ming et al. introduces a new dataset named Real World Fighting (RWF) 2000 which contains 2000 videos captured by surveillance cameras. RWF-2000 is extremely relevant n violence detection because it is based on surveillance videos. It contains 1000 violent and 1000 non-violent videos.

Then we reviewed existing techniques for violence detection or classification of videos into violent and non-violent. Ramzan et al. [2] reviews almost all existing techniques and classifies them into three categories: violence detection techniques using Machine learning, Support Vector Machines and Deep learning. Techniques based on Machine learning and SVMs use feature descriptors such as BoW, MoSIFT, ViF, RIMOC, HOG and HOF, e.t.c. along with SVM, KNN and Adaboost as classifiers. Techniques based on Deep learning rely on CNNs to classify videos. 3D-CNN have been adopted as generally better than 2D-CNN in case of videos. Some methods use a hybrid technique using feature descriptors to extract features and CNNs to classify them.

3. METHODOLOGY

We have followed technique introduced in Ullah et al. [3]. This technique uses three stages to detect violence in videos. First, it uses a MobileNet + SSD to detect humans in videos.

Then it extract 16 frames surrounding the frame containing humans which are fed into a 3D-CNN which classifies them into violent or non-violent frames. If violence is detected signals are sent to nearby authorites to respond.

In our approach, we are not using a MobileNet + SSD (introduced by Howard et al. [4]) to detect humans in frames. The rationale behind it is that RWF-2000 contains only 5 second long vidoes at 30 fps. Also almost all frames contain humans so detecting them would essentially give us whole video back.

We opted the 3D-CNN model. This model itself has been adopted by Ullah et al. from Tran et al. [5] which introduced 3d-CNNs for spatio-temporal feature extraction. 3D-CNN as compared to 2D-CNN preserve temporal information in videos while 2D-CNN only preserves spatial information. Reason is 2D-CNN gives a two dimensional output of a convolution thus compressing the three dimensional input to two. On the other hand, 3D-CNN gives three dimensional output.

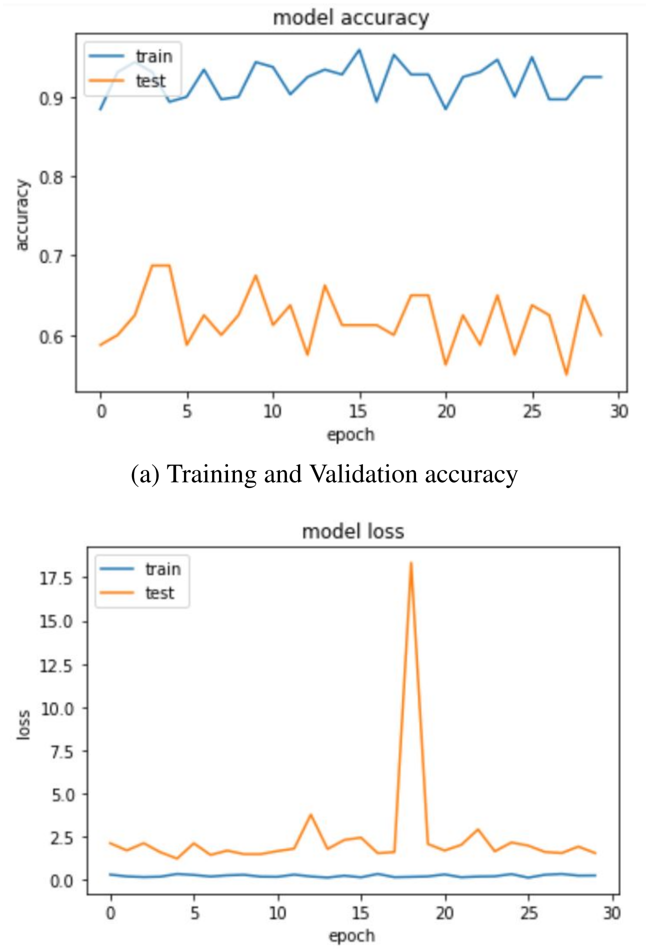
3D-CNN model contains eight convolutional, five max pooling, and two fully connected layers, followed by SoftMax layer (Output). Each convolutional layer has $3 \times 3 \times 3$ kernels with one stride, and all the pooling layers are max pooling with a $2 \times 2 \times 2$ kernel size except for the first pooling layer where kernel size is $1 \times 2 \times 2$ with two strides, preserving the time-based information. Number of filters in each convolution layer increase from 64 in the first layer to 512 in the last layer. Input size of the model is $3 \times 16 \times 112 \times 112$ where 3 and 16 represent RGB channels and number of frames, respectively. Once features have been extracted, Softmax classifier classifies video sequence into violent and non-violent.

To prepare data for our model, we extracted every 9^{th} frame from 150 frames of a video. In this way we reduced number of frames to 16. Also, from 2000 videos (1000 fight and 1000 non-fight) videos, we took a total of 600 videos. For training 320 videos: 160 fight and 160 non-fight. For validation 80 videos: 40 fight and 40 non-fight. For testing 200 videos: 100 fight and 100 non-fight.

4. RESULTS

We ran our model along with the mentioned data setting for 300 epochs. Google Colab GPU environment was used to train model. Obtained results for 270^{th} till 300^{th} epochs have been shown in the figure 1. Training accuracy averaged at about 90% while validation accuracy bounced back and forth 60%. However, testing accuracy was 66% and loss was 2.83.

Flow Gated Model proposed by [1] uses same dataset to train. After 6000 iterations of training on 1400 videos, validating on 200 videos and testing on 400 videos, [1] obtained training accuracy of 95%, validation accuracy of 81% and test accuracy of 86.75%.



(a) Training and Validation accuracy

(b) Training and Validation loss

Fig. 1. Results obtained after training model for epochs 270-300

However, our intuition is that since our model is deeper than the Flow Gated Model, training our model on full dataset and for more epochs, e.g. 500 epochs, would make our accuracy better than Flow Gated Model.

5. CONCLUSION

We conclude our project with idea that large dataset, which is close to real surveillance data, and deeper model will drastically improve performance of violence detection deep learning models. For future work, we suggest developing an algorithm which extracts least correlated frames from a video instead of just picking every 9^{th} frame. Other possible future works are making a hybrid model which extracts features from frames using video descriptors and combines them with the results of 3D-CNN.

6. REFERENCES

[1] Cheng M, Cai K, and Li M, “Rwf-2000: An open large scale video database for violence detection,” Nov 2019.

[2] Ramzan M et al., “A review on state-of-the-art violence detection techniques,” *IEEE Access*, vol. 7, pp. 107560–107575, 2019.

[3] Ullah F, Ullah A, Muhammad K, Haq I, and Baik S, “Violence detection using spatiotemporal features with 3d convolutional neural network,” *Sensors*, May 2019.

[4] Howard G et al., “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” april 2017.

[5] Tran D, Bourdev L, Fergus R, Torresani L, and Paluri M, “Learning spatiotemporal features with 3d convolutional networks,” Oct 2015.