

Automatic video censoring system using deep learning

Yash Verma¹, Madhulika Bhatia², Poonam Tanwar³, Shaveta Bhatia⁴, Mridula Batra⁴

¹Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University, Noida, India

²Department of Computer Science and Engineering, Manav Rachna International Institute of Research and Studies, Faridabad, India

³Faculty of Engineering and Technology, Manav Rachna International Institute of Research and Studies, Faridabad, India

⁴Faculty of Computer Applications, Manav Rachna International Institute of Research and Studies, Faridabad, India

Article Info

Article history:

Received Jun 9, 2021

Revised Jun 21, 2022

Accepted Jul 18, 2022

Keywords:

Automatic video censoring

Computer vision

Convolutional neural network

Deep learning

Machine learning

ABSTRACT

Due to the extensive use of video-sharing platforms and services, the amount of such all kinds of content on the web has become massive. This abundance of information is a problem controlling the kind of content that may be present in such a video. More than telling if the content is suitable for children and sensitive people or not, figuring it out is also important what parts of it contains such content, for preserving parts that would be discarded in a simple broad analysis. To tackle this problem, a comparison was done for popular image deep learning models: MobileNetV2, Xception model, InceptionV3, VGG16, VGG19, ResNet101 and ResNet50 to seek the one that is most suitable for the required application. Also, a system is developed that would automatically censor inappropriate content such as violent scenes with the help of deep learning. The system uses a transfer learning mechanism using the VGG16 model. The experiments suggested that the model showed excellent performance for the automatic censoring application that could also be used in other similar applications.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Madhulika Bhatia

Department of Computer Science and Engineering, Amity School of Engineering and Technology

Amity University

Amity Rd, Sector 125, Noida, Uttar Pradesh 201301, India

Email: madhulikabhatia@gmail.com

1. INTRODUCTION

“We become what we see” is an ancient famous quote that holds even today. It points out the importance of the visual information effect on people in a way that is profoundly ingrained. Internet is full of things that are inappropriate for children and even teenagers that can affect their young minds in a very negative way. Violent scenes that encourage young viewers to reproduce the roles performed on television by their characters will cost them their lives. Films claiming that drug use is health-destroying simply promote drug use as young minds become curious to do it in real life. These scenes should be censored tighter.

The internet's media presence is vast because of the widespread use of video sharing sites and cloud facilities. This data volume makes it impossible to monitor the content of those videos. One of the most crucial issues about the video contents is whether it has an objectionable topic, for example, violence or abuse in any form. More than telling if a video is either appropriate or inappropriate, it is also important to identify which parts of it contain such content, for preserving parts that would be discarded in a simple broad analysis.

Censoring boards and committees are there for large scale movies and television shows, but people are moving to other sources like smartphones, tablets, and computers. as their primary source of entertainment from movies theatres and Television shows. There are several sources of information and

entertainment like Netflix, Amazon prime video, YouTube, Instagram, Facebook, and Twitter. As these can be very convenient at times due to their portability and reach to the major population in the world. A mechanism has not been put into place to censor the content which gets shared through these platforms. Especially in developing countries like India, this is a big issue. As opposed to the primary movies given by central board of film certification (CBFC), the broadcasting content complaints council (BCCC), and the over-the-top (OTT) platforms do not have a regulatory authority to monitor the content streamed and therefore enjoy its rights. The contents on these websites are also subject to observation by the supreme court direct contradiction to the various laws of the country. But the legal loopholes and grey fields are also troubling [1]. Also, many notorious groups deliberately put inappropriate content on social media. There were many instances where people put up content related to unbearable violence and bloodshed that is unbearable even for fully grown adults. The proposed system will help in filtering the above-said content.

Censoring content in the videos may be specified as an example of binary classification where two different classes can be appropriate and non-appropriate. Some of the popular machine learning (ML) are logistic regression, k-nearest neighbors, decision trees, support vector machines (SVM) and naïve-Bayes algorithm. But here the data to be classified is in the form of complex images that vary very much. Due to this complex data, traditional ML algorithms are not enough as discussed in the literature review section below. So, this problem has to be tackled by deep learning (DL). The best known and most effective deep learning technique is the convolutional neural network. Through this work, different deep learning models have been compared to find the model that is best suited for the task of automatic censoring of videos and images by classifying the frames as violent or non-violent. And the complete architecture of the system is provided in this paper along with the results and at last, conclusions are drawn.

A lot of research was conducted for the domain of video censoring using machine learning and deep learning. Some of the most notable work relevant to this project has been mentioned. Nievas *et al.* [2] evaluated the various state of the art video descriptors for the task of detection of fights on datasets containing action movie scenes and National Hockey League footage. The popular bag of words approach has been found to 90% accurately classify the frames. It was discovered that for hockey data set that the accuracy of the choice of feature descriptor of low level and vocabulary size is not important; furthermore, on the latter dataset, descriptor choice was critical; under either circumstance, the motion SIFT (MoSIFT) functionality was significantly greater than the best spatio-temporal interest points (STIP).

Deniz *et al.* [3] proposed a new approach for the detection of violent actions in which the major discriminatory attribute is extreme acceleration patterns. The proposed algorithms resulted in 12% accuracy improvement over state-of-the-art methods in surveillance scenarios including sports where complex actions take place. It was hypothesized from the experiments that motion alone is sufficient for categorization and visuals could be a cause of confusion in the detection process and cost any additional computations. The method found to be 15-fold faster than with a very a smaller number of features. This can also be used as the first stage of a maximum accuracy system with STIP or MoSIFT features.

Fu *et al.* [4] presented an efficient approach for violence detection in videos using analysis of motion without action events, gestures, or complex behavior recognition. A heuristic framework has been suggested, built on a decision tree structure, to derive more accurate motion information from optical flow images to distinguish movement types by motion, count, size, and direction according to the movement regions extract. Motion attraction was also suggested to measure intensity between two motion zones, which will measure different statistics as classification features.

Bilinski and Bremond [5] proposed a technique that is based on improved fisher vectors (IFV) allowing for both spatio-temporal as well as local features for video representation for the purpose of recognizing violence in videos. When comparing the proposed technique with temporal-spatial positions IFV, the proposed extension got similar or better accuracy. This new strategy has been more effective than the previous approaches for violence detection too. Here, the approach using sliding-window have also been studied and IFV has been reformulated for increasing the speed of the framework for detection of violence on four states of the art datasets.

Nar *et al.* [6] showed that for abnormal behavior detection in front of automatic teller machines (ATMs), the recognition of posture could be used. The experiments were done on the data captured using Kinect 3D camera. Logistic regression was used as the classification technique. For the calculation of optimal parameters, gradient descent was used.

Xie *et al.* [7] proposed an algorithm using a motion vector for the detection of violence for surveillance videos. First, the system removes motion vectors from compressed videos and next analyses the space-time distribution of the magnitude of the motion vectors and the path to take region motion vectors (RMV), then uses radial based SVM, which classifies RMV and determines violent behavior in monitoring videos. It can increase video-monitoring systems performance on the detection of violent activities in real-time and increase the retrieval systems performance for videos on the position of the violence in historical videotapes or other media source. The methodology provided is highly appropriate for front video

encoders operating on an integrated digital signal processor (DSP) network since it saves measures to identify and trace movements that are normally applied in the conventional method of behavior analysis. The provided methodology could classify the UCF sports dataset with 91.9% accuracy.

Ribeiro *et al.* [8] worked on the issue of detection of violence where detection is difficult due to some external factors like backgrounds that are clustered and dynamic, occlusion making the scene complex for detection. A proposition was given for rotation-invariant feature modelling motion coherence that is specific to the violence detection problem. That is used to differentiate against unstructured movements and structure capture. The value of the histogram of optical flow vectors, obtained from the 2nd order statistics of time instants, is calculated locally, densely, and integrated into the Riemannian spherical manifold. It is dependent on its values. It was shown through experiments, the accuracy given by the provided approach similar to those of state-of-the-art models in the laboratory as well as real-world setting.

Senst *et al.* [9] presented an automatic special violence detection technique based on the Lagrangian technique. We propose a new feature based on a spatio-temporal model, which uses appearance, context motion compensation, and details on the longer-term motion. We provide an extensive bag-of-words procedure as a per-video classification scheme to ensure suitable spatial and temporal scales. The proposed architecture was thoroughly reviewed by the London metropolitan police on multiple challenge data sets and real-world non-public data.

Zhou *et al.* [10] proposed FightNet, convolutional neural network (CNN) for modelling long-term temporal structure for detecting interactions that are violent. The model was trained on the UCF101 dataset. Acceleration fields are researched for capturing motion features and found to be responsible for a 0.3% increase in inaccuracy. The system was able to achieve higher accuracy with decreased computational cost. Firstly, the framing of each video is red, green, blue (RGB). Second, the field of optical flow is calculated by following photos. Acceleration is obtained by the optical flow field. Third, FightNet has three types of input modes, including RGB images, optical flow images and acceleration images for temporal networks. When fused from all inputs, we infer whether a video says a violent incident or not.

Chaudhary *et al.* [11] put forward a technique for automatically detecting violence in surveillance videos. The method proposed contains three key steps: moving object identification, observing objects and comprehension of behavior for movement recognition. Key features (speed, direction, center, and dimensions) are defined by using the feature extraction method. This helps trace objects in video frames. For extracting foreground, the Gaussian mixture model was utilized. This method could categorize the videos with 90% accuracy.

Zhou *et al.* [12] also proposed a different algorithm involving optical flow fields. First, the movement regions are divided by optical flow field distribution. Next, in motion areas, a proposal was made that the presence and dynamics of aggressive actions be extracted from two kinds of low-level features. The suggested low-level characteristics of the local appendix histogram (LHOG), derived from the local appendix optical flow histogram (LHOF), extracted from optic flow photographs, are the Low-level characteristics. Now, for each video, a vector of a certain length is obtained, and the features collected are coded using the bag-of-words (BoW) to remove unnecessary content. The last thing was to use SVMs to classify vectors at the level of each video.

Khan *et al.* [13] proposed a model to classify cartoon content as inappropriate for children. The model used the transfer learning approach taking benefit of the MobileNet model. The system proved to be 97% accurate.

Song *et al.* [14] made a proposal on 3D ConvNets with modified pre-processing steps for the detection of violence. This paper proposes a new sampling method in which instead of uniform sampling, keyframes are identified and those are implemented for categorization of sequence instead of every frame. Through this method, as well as exceptional precision on benchmark datasets, the computing is greatly reduced which result in faster classification. An important point is that the shorter clips are treated in a similar manner as uniform sampling rate, just lengthy clips use the new method which further maintains the accuracy and speed. For three public violent detection datasets: hockey fight, movies, and crowd violence, individualized strategies are implemented to suit the varied clip length. They used uniform sampling for short clips. However, for longer videos, the fixed sample method brings the problem of redundancy and the discontinuity of motions. A new method is proposed for longer clips. The proposed scheme obtains competitive results: 99.62% on hockey fight, 99.97% on movies, and 94.3% on crowd violence.

Khan *et al.* [15] proposed a model based on a similar approach as the Song *et al.* [14]. Initially, the entire film is divided into images, and then an individual image is chosen from each scene depending on saliency intensity. These pictures can then be translated from a lightweight model, which is well tuned to distinguish aggression and nonviolence in a film using a transfer learning technique. Finally, all nonviolent scenes are combined into a series to provide a non-violent film that children can enjoy, and paranoid persons can watch. The model is tested on benchmark datasets and good accuracy has been achieved.

Freitas *et al.* [16] proposed a multi-model method that classifies inappropriate and appropriate scenes using both visual and audio parameters using convolutional neural networks. The InceptionV3 has been used for video and AudioVGG for audio classification. Then the principal component analysis is used for feature selection and finally, SVM performs the last categorization. The model achieved 98.94% and 98.95% F1 scores for inappropriate and appropriate content, respectively.

Gkountakos *et al.* [17] proposed a framework for the detection of violence by crowd-based using ResNet and 3D ConvNets. The framework is compared with several states of the art models on the Violent-Flow dataset. The 3D-ResNet50 framework proved to give better accuracy and is quicker than the compared models from the experiments using Violent-Flows dataset.

Roman and Chávez [18] proposed an approach for violence detection and localization in surveillance videos. The proposed method is based on ConvNets and dynamic images. Instead of using the computationally costly optical flow, researchers used dynamic images which besides reducing the cost of computation also helped the analysis of long temporal information. This helped in getting good accuracies with lower computational cost.

Li *et al.* [19] proposed a multiple stream method for violence detection for the video surveillance system. The solution suggested improves the detection of acts of violence in the video by merging three separate streams: spatial RGB source, time stream and local space stream. The focus-based spatial RGB stream discovers from soft-attention mechanisms the spatial attention regions of people that are highly likely to be action regions. As the input to retrieve temporal characteristics, the temporary stream uses optical flow. The local spatial stream uses block images as feedback to learn spatial local characteristics. The algorithm's proposal was tested on a self-compiled elevator surveillance dataset and found to be satisfactory.

Accattoli *et al.* [20] used C3D, which is a 3D ConvNet architecture that allows the computation of video descriptor features to remove motion features without any previous information. These descriptors were then used to characterize videos as either aggressive or peaceful bypassing descriptors as an input for a linear support vector machine. For the model, performance was improved for the application of both crowd and individual violent action recognition than the state-of-the-art models.

The technique developed by Sharma *et al.* [21] was composed of three phases. In the first step, the whole film is split into shots, and then a random sample from each shot is chosen. The next step is to transfer learning, which was done to fine-tune the classification of violence and nonviolence. To put an end to all the non-violence sequences, the violence-free movie that can be seen by children and violent paranoid individuals is made by stitching all the non-violence segments together. The model was evaluated on the Violence in movies dataset, the Hockey fights dataset, and the VSD dataset benchmarks, and it obtained an overall accuracy of 96.3%. MobileNetV2 [22] has been a popular lightweight model for the task of object detection. Xception [23], InceptionV3 [24], VGG16 [25], ResNet50 [26] are some of the winners for the ImageNet challenges in different years that showed excellent accuracies for the ImageNet dataset.

2. METHOD

2.1. Dataset

The movies fight detection dataset [27] has been used for comparison purpose. The dataset is comprised of 200 videos of variable length taken from various international movies. For computation, the individual frame images have been extracted from these videos. A few samples of nonviolent and violent images from the data set have been provided as shown in Figures 1 and 2.



Figure 1. Non-violent images from the dataset

2.2. Hardware and software used

The implementation was done on a computer with an CPU Intel(R) Core(TM) i7-9750H CPU @ 2.60 GHz, GPU NVIDIA GeForce GTX 1660 Ti, RAM 8 GB, and operating system Windows 10. This research used language Python 3.7.10. And used Framework Keras with TensorFlow 2.4.1.



Figure 2. Violent images from the dataset

2.3. System methodology

The overall methodology is explained using the following steps. The complete workflow of the system has been illustrated through the flowchart in Figure 3. Input raw video is taken from dataset. The frames were splits from the uploaded video. Image processing techniques applied using deep neural network. The pretrained dataset also been fed into the model. The videos were classified into violent or not using pass unchanged and gaussian blurring.

2.3.1. Input raw video

The first step is the input step. The system can accept both image and video input. The system can be in one of the common formats for images like JPG or PNG and mp4 for videos. The minimum expected resolution of the input is 240×240 pixels.

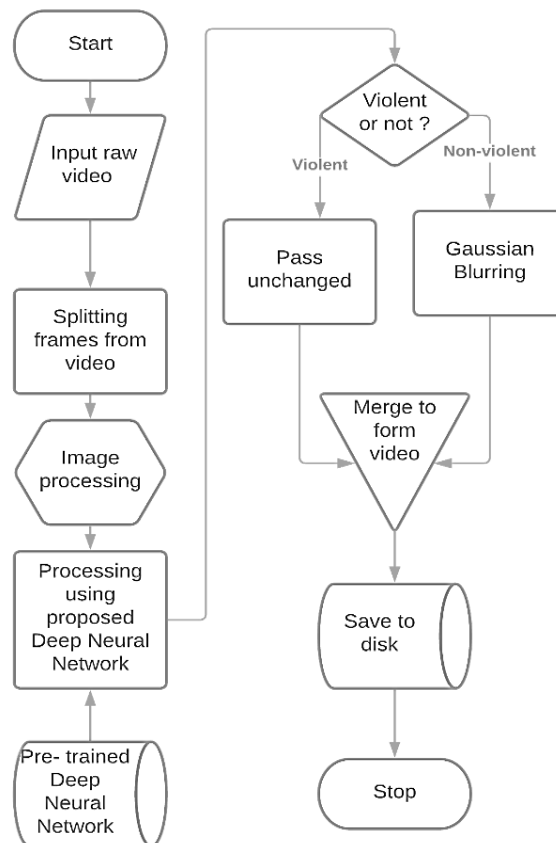


Figure 3. Methodology flowchart

2.3.2. Splitting frames from video

Every video consists of a series of frames. So, if the input is in the form of a video, it is first separated into individual frames. The system is capable of handling a high frame rate but for testing purposes

to decrease the processing time (because of unavailability of high specification hardware), the sampling rate here has been set to ten frames per second.

2.3.3. Image processing

To give more flexibility input could be in any resolution greater than 240×240 pixels. But the system is designed to work upon input of 240×240 pixels, all the frames are converted to this dimension. This resolution has been chosen owing to the hardware bottleneck. So, in presence of hardware of higher specifications, the system can be easily adjusted to work upon higher resolution frames.

2.3.4. Processing using proposed deep neural network

The processed frames are fed to the DL model as illustrated below and the output is one of two categories i.e., violent, or non-violent. There are techniques which are used for processing of videos using deep neural network are VGG19, Convolution 2D layer as well as Max pool layer. The architecture of the model is explained below and illustrated in Figure 4.

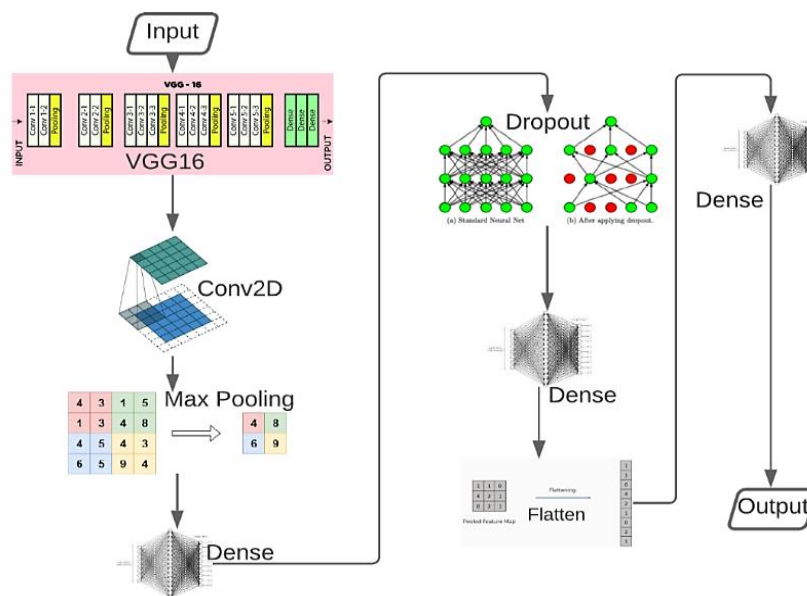


Figure 4. Proposed deep neural network architecture

– VGG16

The model utilized the pre-trained model weights to save time and other resources. VGG16 is an excellent object detection model. So, the image input is first passed to VGG16, and the output of the second last layer is passed to the convolutional 2D (Conv2D) layer of the new model instead of the output layer.

– Convolutional 2D layer

Conv2D detects various features in frames via convolution operation. A kernel is passed over the whole image or feature map and the result of convolution is stored in the output feature map. The features from frames were extracted to classify video content as violent or nonviolent. These frames were further combined and saved in disk.

– Max pooling layer

Max pooling is one of the three pooling techniques used for dimensionality reduction in CNNs. In the max-pooling layer, the maximum value of the feature is selected from specified adjacent values of feature maps. The feature maps received certain values which is useful in selecting the highest values.

– Dense layer

In this layer, every neuron from the previous layer is connected to every neuron. That is why it is also called a fully connected layer. Matrix multiplication takes place in this layer in the background and learning takes place via adjustment of weights by backpropagation mechanism.

– Dropout layer

The introduction of this layer is important to reduce the model overfitting while training the model employing regularization. A dropout factor of 0.3 has been set here which means 30% of the inputs are

randomly discarded and others are moved to dense layer. The dense layer is created which is further used to accumulate those portions of inputs which has thirty% dropout.

– Dense layer

This layer function in a similar way as the above dense layer to enforce learning in the model after the dropout layer. This is the last layer processing multi-dimensional inputs and outputs. This layer is deeply connected neural network layer is very common and frequently used layer. Dense layer operates on frames of input and return the output.

– Flatten layer

Flatten layer is used to convert multidimensional feature maps into a long single-dimensional vector so that the final classification can be performed by the fully connected output layer quickly. It makes the multidimensional frame input one-dimensional frame output. This layer helpful in transition of convolution layer to the fully connected layer.

– Dense output layer

This is the last layer in the proposed model where actual classification takes place. Here, the sigmoid activation function is used for classification. The output of this layer is one of the two classes that is violent or non-violent.

2.3.5. Applying blur to violent images

Based on the output of the model, gaussian blur is applied to the frame. If the frame is violent, the frame is blurred otherwise no change is made to the frame. The output is recognized as violent and differentiated with nonviolent frames on the basis of blurred texture observed in the frame.

2.3.6. Merge to form video

The processing till now is done on individual frames. In this step, all the frames are again combined and encoded to form the video output (or image in case of image inputs). The multidimensional videos are converted into one dimensional. The unit frames merged and encoding is performed to get the video back.

2.3.7. Saving to disk

In this step, the output video or image file is stored on the storage medium at the path specified that is the same path as the input file by default. After classifying video into violent and non-violent videos. These videos will be merged and stored in local hard disk.

3. RESULTS AND DISCUSSION

It has been inferred from the literature review that the deep learning approach has been found more effective than the alternative ML approaches. So, some of the popular deep learning models are assessed for the application of classifying violent and non-violent scenes. For comparison, the last layer of every network compared was removed and the base model layers have been set as non-trainable. All the models used for transfer learning have been concatenated with similar layer architecture so that the results will not be biased due to anything except base model architecture. Tables 1 to 4 show the values of training accuracies, validation accuracies, training losses, and validation losses respectively for all the considered models for the same dataset.

Table 1. Training accuracies

Epoch	VGG16	ResNet50	ResNet101	Xception	InceptionV3	MobileNetV2	VGG19
1	0.4308	0.5325	0.4286	0.6099	0.4009	0.3329	0.4433
2	0.6145	0.6699	0.5435	0.7211	0.4043	0.3607	0.5779
3	0.6286	0.7788	0.5546	0.7373	0.4378	0.3874	0.5627
4	0.6259	0.6919	0.5710	0.7079	0.4533	0.3342	0.5877
5	0.6259	0.7120	0.5736	0.7673	0.4715	0.3936	0.5976
6	0.6708	0.7923	0.6016	0.7770	0.5085	0.4334	0.6051
7	0.6708	0.7993	0.6174	0.7452	0.5264	0.4489	0.5783
8	0.7814	0.8766	0.5928	0.7200	0.5438	0.4456	0.5879
9	0.8476	0.9482	0.5917	0.7800	0.6046	0.4433	0.6116
10	0.9521	0.7746	0.5958	0.7771	0.5867	0.4530	0.7018
11	0.9743	0.8658	-	-	-	-	-
12	0.9962	0.9887	-	-	-	-	-
13	0.9968	0.9969	-	-	-	-	-
14	0.9984	0.9963	-	-	-	-	-
15	0.9971	0.9962	-	-	-	-	-

Table 2. Validation accuracies

Epoch	VGG16	ResNet50	ResNet101	Xception	InceptionV3	MobileNetV2	VGG19
1	0.3811	0.4325	0.1022	0.2596	0.0899	0.0688	0.1527
2	0.3802	0.6518	0.1178	0.2362	0.0899	0.0651	0.1633
3	0.3848	0.5495	0.1165	0.2495	0.0899	0.0894	0.1458
4	0.3926	0.3527	0.1220	0.2614	0.0912	0.0880	0.2183
5	0.2472	0.3100	0.1293	0.2665	0.0958	0.0889	0.1853
6	0.2495	0.3729	0.2004	0.2596	0.1073	0.0894	0.272
7	0.5504	0.3743	0.2266	0.2160	0.1178	0.0894	0.2192
8	0.6307	0.4752	0.2243	0.2798	0.1688	0.0889	0.2087
9	0.7834	0.4325	0.2220	0.2807	0.2137	0.0889	0.2692
10	0.8853	0.5330	0.1422	0.3036	0.3582	0.0889	0.2917
11	0.9706	0.7853	-	-	-	-	-
12	0.9486	0.8426	-	-	-	-	-
13	0.9559	0.9220	-	-	-	-	-
14	0.9559	0.7261	-	-	-	-	-
15	0.9522	0.7775	-	-	-	-	-

Table 3. Training losses

Epoch	VGG16	ResNet50	ResNet101	Xception	InceptionV3	MobileNetV2	VGG19
1	0.0846	0.0974	0.6127	3.9513	2.0313	7.8345e-05	0.2149
2	0.0020	0.0023	0.0002	0.0255	0.3080	7.4334e-05	0.0002
3	2.1427e-05	0.0027	0.0013	0.6266	0.2705	0.0002	0.0004
4	7.4089e-05	0.0108	7.5943	0.0084	0.2483	0.0013	0.0021
5	0.0009	0.0043	0.0002	0.0021	0.2207	2.8252e-05	3.23E-05
6	6.3740e-06	0.0001	0.0004	0.0014	0.1909	1.8697e-05	0.0002
7	0.0023	5.2149	0.0017	0.0194	0.1526	2.3616e-05	6.62E-05
8	0.0116	0.0010	2.1748	0.0053	0.1409	8.5036e-06	2.65E-05
9	0.0200	0.0054	3.5433	0.0057	0.0953	1.9839e-05	3.90E-03
10	0.0169	0.0001	0.0008	0.0010	0.0884	2.2081e-05	2.50E-03
11	0.2106	0.0566	-	-	-	-	-
12	0.1233	0.0756	-	-	-	-	-
13	0.0009	0.0361	-	-	-	-	-
14	1.6412e-12	0.1541	-	-	-	-	-
15	0.0088	0.2933	-	-	-	-	-

Table 4. Validation losses

Epoch	VGG16	ResNet50	ResNet101	Xception	Inception	MobileNet	VGG19
1	0.1721	0.2322	0.9444	8.4396	1.1195	2.7507	0.5384
2	0.5360	0.2199	1.1235	8.1314	2.6441	3.0738	0.2601
3	0.4174	0.0651	1.3611	9.5290	5.0485	1.4225	0.7005
4	0.349	0.2968	1.6463	15.6250	13.4026	2.5173	0.1774
5	0.9826	1.9352	2.5149	16.8107	6.4081	2.8827	0.3428
6	1.1410	1.5655	3.1745	11.7113	13.0884	2.8260	0.1015
7	4.2067	1.5548	0.4659	6.7274	21.7036	2.7518	0.1844
8	1.8556	4.4436	0.4944	8.6459	5.3421	2.9729	0.2076
9	1.4086	2.1964	0.5038	13.4710	13.4546	2.8615	4.56
10	0.5593	2.4719	1.2660	9.0726	2.1627	3.2106	1.1646
11	0.1909	1.9767	-	-	-	-	-
12	0.4748	2.9610	-	-	-	-	-
13	1.1006	3.2689	-	-	-	-	-
14	1.1067	25.5439	-	-	-	-	-
15	4.8550	6.9018	-	-	-	-	-

Shown below are the model training and validation metrics. Both the models have shown considerably lower training losses than the other compared models. But the validation losses in the case of VGG16 have been found to be significantly lower when compared to the validation losses when using ResNet50.

The time required for validation and training of ResNet101 and VGG19 is more than double as of ResNet50 and VGG16 as being the most complex model among the compared models, but the accuracy is worse than the ResNet50 and VGG16. Xception lies somewhere between VGG16 and ResNet101. The MobileNetV2 being the lightest model has also got the least training time, but the accuracy is also worse than VGG16 and ResNet50. The training time for VGG16, ResNet50 and Inception are similar.

Since VGG16 and ResNet50 have shown the best accuracies among all, they are trained for 5 more epochs to get the most accurate models. Both the models start to overfit after epoch 13. Also, the validation and training losses are at optimum at epoch 13, as observed from Figures 5 to 12. So, the model training is

stopped at epoch 13 at 99.68% training and 95.59% validation accuracy for VGG16, and 99.69% training and 92.20% validation accuracy for ResNet50.

The validation and training losses are also at optimum at epoch 13, as observed from Figures 6, 8, 10 and 12. So, the model training is stopped at epoch 13 at 99.68% training and 95.59% validation accuracy for VGG16, and 99.69% training and 92.20% validation accuracy for ResNet50. The VGG16 performed well in respect of accuracy as 99% as compared to validation accuracy achieved by ResNet.

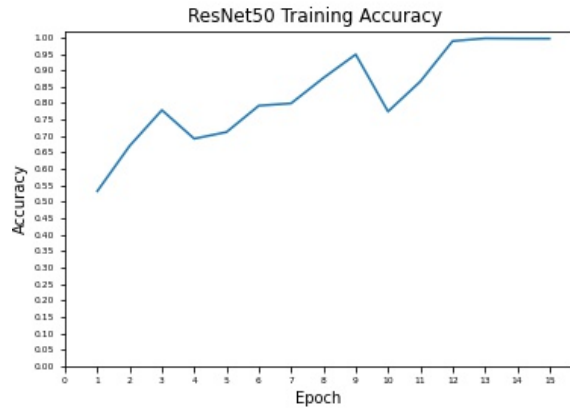


Figure 5. ResNet50 training accuracy

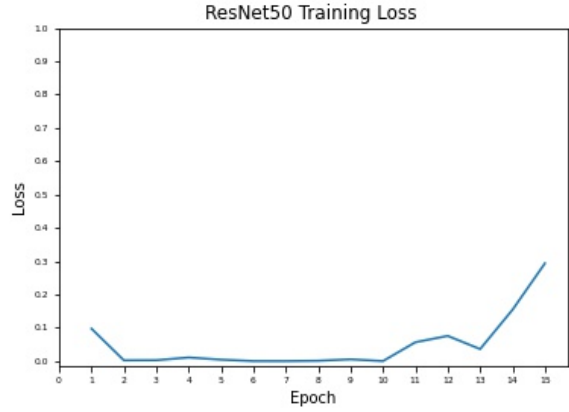


Figure 6. ResNet50 training loss

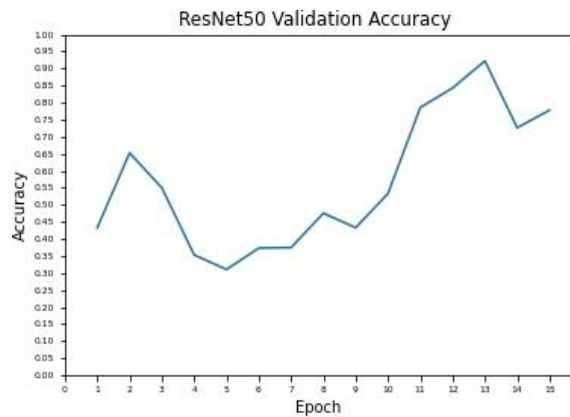


Figure 7. ResNet50 validation accuracy

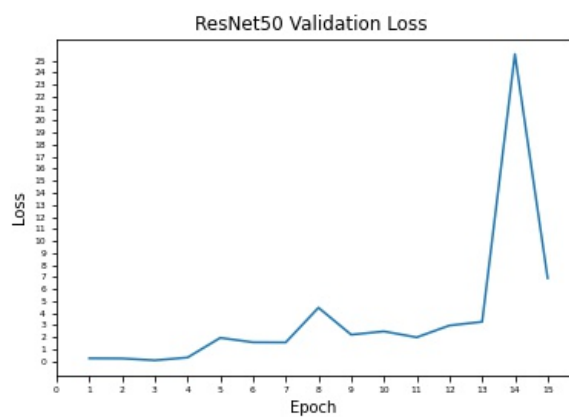


Figure 8. ResNet50 validation loss

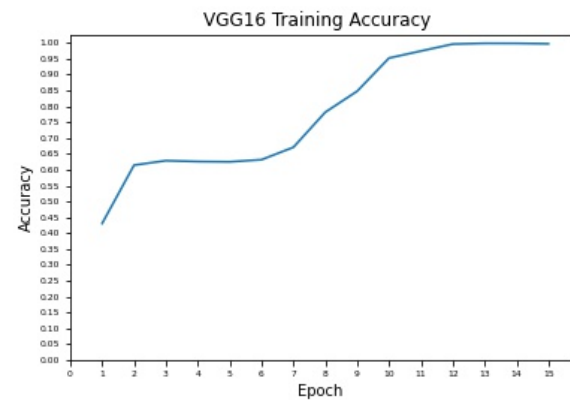


Figure 9. VGG16 training accuracy

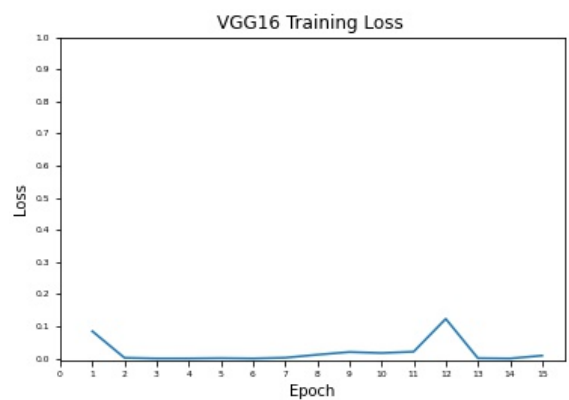


Figure 10. VGG16 training loss

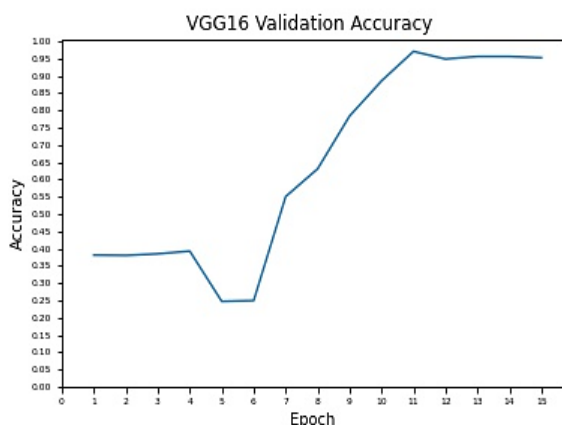


Figure 11. VGG16 validation accuracy

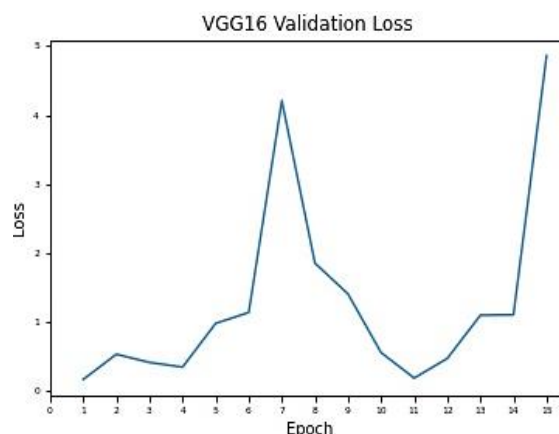


Figure 12. VGG16 validation loss

4. CONCLUSION

VGG16 and ResNet50 has shown best results among MobileNetV2, Xception model, InceptionV3, ResNet101, ResNet50 and VGG16. It has been concluded from the experiments that VGG16 is the most appropriate model for designing the system to classify the violent and non-violent scenes from the movies. This is shown to be 99.68% for training and 95.59% accurate for validation. Due to the absence of hardware with high specifications, the system is made to reduce the resolution of output to 240×240 pixels and the sampling rate is reduced to 10 frames per second. The system is capable to work at high sampling rates and work for high-resolution output if better hardware is used for processing.




REFERENCES

- [1] Nagoriastha, "Censorship of OTT platforms: a boon or bane," Legal Service India. <http://www.legalserviceindia.com/legal/article-3418-censorship-of-ott-platforms-a-boon-or-bane.html> (accessed May 20, 2021).
- [2] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthakar, "Violence detection in video using computer vision techniques," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6855, no. 2, Springer Berlin Heidelberg, 2011, pp. 332–339, doi: 10.1007/978-3-642-23678-5_39.
- [3] O. Deniz, I. Serrano, G. Bueno, and T. K. Kim, "Fast violence detection in video," in *9th International Conference on Computer Vision Theory and Applications*, 2014, vol. 2, pp. 478–485, doi: 10.5220/0004695104780485.
- [4] E. Y. Fu, H. Va Leong, G. Ngai, and S. Chan, "Automatic fight detection in surveillance videos," in *14th International Conference on Mobile Computing and Multi Media*, Nov. 2016, pp. 225–234, doi: 10.1145/3007120.3007129.
- [5] P. Bilinski and F. Bremond, "Human violence recognition and detection in surveillance videos," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug. 2016, pp. 30–36, doi: 10.1109/AVSS.2016.7738019.
- [6] R. Nar, A. Singal, and P. Kumar, "Abnormal activity detection for bank ATM surveillance," in *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sep. 2016, pp. 2042–2046, doi: 10.1109/ICACCI.2016.7732351.
- [7] J. Xie, W. Yan, C. Mu, T. Liu, P. Li, and S. Yan, "Recognizing violent activity without decoding video streams," *Optik*, vol. 127, no. 2, pp. 795–801, Jan. 2016, doi: 10.1016/j.ijleo.2015.10.165.
- [8] P. C. Ribeiro, R. Audigier, and Q. C. Pham, "RIMOC, a feature to discriminate unstructured motions: application to violence detection for video-surveillance," *Computer Vision and Image Understanding*, vol. 144, pp. 121–143, Mar. 2016, doi: 10.1016/j.cviu.2015.11.001.
- [9] T. Senst, V. Eiselein, A. Kuhn, and T. Sikora, "Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 2945–2956, Dec. 2017, doi: 10.1109/TIFS.2017.2725820.
- [10] 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, no. 1, Jun. 2017, doi: 10.1088/1742-6596/844/1/012044.
- [11] S. Chaudhary, M. A. Khan, and C. Bhatnagar, "Multiple anomalous activity detection in videos," *Procedia Computer Science*, vol. 125, pp. 336–345, 2018, doi: 10.1016/j.procs.2017.12.045.
- [12] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violence detection in surveillance video using low-level features," *PLOS ONE*, vol. 13, no. 10, Oct. 2018, doi: 10.1371/journal.pone.0203668.
- [13] N. F. Khan, A. Hussain, A. Khan, and I. U. Din, "Categorized violent contents detection in cartoonmovies using deep learning model: mobile net," in *5th International Conference on Next Generation Computing*, 2019.
- [14] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A novel violent video detection scheme based on modified 3D convolutional neural networks," *IEEE Access*, vol. 7, pp. 39172–39179, 2019, doi: 10.1109/ACCESS.2019.2906275.
- [15] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, "Cover the violence: a novel deep-learning-based approach towards violence-detection in movies," *Applied Sciences*, vol. 9, no. 22, Nov. 2019, doi: 10.3390/app9224963.
- [16] P. V. A. de Freitas et al., "A multimodal CNN-based tool to censure inappropriate video scenes," *arXiv:1911.03974*, Nov. 2019.
- [17] K. Gkountakos, K. Ioannidis, T. Tsirikra, S. Vrochidis, and I. Kompatsiaris, "A crowd analysis framework for detecting violence scenes," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, Jun. 2020, pp. 276–280, doi: 10.1145/3372278.3390725.




- [18] D. G. C. Roman and G. C. Chavez, "Violence detection and localization in surveillance video," in *33rd SIBGRAPI Conference on Graphics, Patterns and Images*, Nov. 2020, pp. 248–255, doi: 10.1109/SIBGRAPI51738.2020.00041.
- [19] H. Li, J. Wang, J. Han, J. Zhang, Y. Yang, and Y. Zhao, "A novel multi-stream method for violent interaction detection using deep learning," *Measurement and Control*, vol. 53, no. 5–6, pp. 796–806, May 2020, doi: 10.1177/0020294020902788.
- [20] S. Accattoli, P. Sernani, N. Falcionelli, D. N. Mekuria, and A. F. Dragoni, "Violence detection in videos by combining 3D convolutional neural networks and support vector machines," *Applied Artificial Intelligence*, vol. 34, no. 4, pp. 329–344, Mar. 2020, doi: 10.1080/08839514.2020.1723876.
- [21] K. Sharma and M. Bhatia, "Deep learning in pandemic states: Portrayal," *International Journal on Emerging Technologies*, vol. 11, no. 3, pp. 462–467, 2020.
- [22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.
- [23] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.
- [24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
- [25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, Sep. 2014, doi: 10.48550/arXiv.1409.1556.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [27] I. S. Gracia, O. D. Suarez, G. B. Garcia, and T. K. Kim, "Fast fight detection," *PLoS ONE*, vol. 10, no. 4, Apr. 2015, doi: 10.1371/journal.pone.0120448.

BIOGRAPHIES OF AUTHORS






Yash Verma    is a student pursuing Master of Technology degree in Computer Science and Engineering from Amity University, Noida, Uttar Pradesh, India. He completed his Bachelor of Technology degree in Computer Science and Engineering from SRM Institute of Science and Technology, Chennai, Tamil Nadu, India. His primary interest is in "Image and Video processing using Machine Learning and Deep Learning Techniques". His research is focused on Security and Surveillance and automatic censorship in videos. He has also worked on a system to train ground security personnel for surveillance in different environments which is compatible with the new age Virtual Reality systems. He can be contacted at email: yashver96@gmail.com.






Madhulika Bhatia    is working as an Associate Professor in the Department of Computer Science and Engineering at Amity School of Engineering and Technology, Amity University, Noida. She holds a Diploma in Computer Science and Engineering, B.E in Computer Science and Engineering, MBA in Information Technology, M. Tech in Computer Science and PhD from Amity University, Noida. She has a total of 15 years of teaching experience. She published almost 32 Research Papers in National, International conferences and Journals. She is also the author of two books. She Filed two Provisional Patent. She attended and organized many workshops, Guest Lectures, seminars. She is also a member of many technical societies like IET, ACM, UACEE, IEEE She reviewed for Elsevier-Heliyon, IGI, Indian Journal of Science and Technology and as well as did Editorial for Springer Nature, Switzerland for Book Chapter in Data Visualization and Knowledge Engineering. She has guided 5 M.Tech. Thesis and around 50 B. Tech. Major and Minor Projects and guiding PhD scholars. He can be contacted at email: madhulikabahtia@gmail.com.






Poonam Tanwar    has 18 years of Teaching Experience working as Prof. in Manav Rachna International Institute of Research and Studies, Faridabad. She has filled 6 patents. She was Guest Editor for Special issue of "Advancement in machine learning (ML) and Knowledge Mining (KM)" for International Journal of Recent Patents in Engineering (UAE). She has organized various Science and Technology awareness program for rural development. Beside this she has even published more than 40 research papers in various International Journals and Conferences. 3 edited books with international publisher are in her credit. She is technical program committee member for various International Conferences. She can be contacted at email: poonamtanwar.fet@mrii.edu.in.



Shaveta Bhatia    is a professor at Department of Computer Applications, MRIIRS. Deputy Director, Manav Rachna Online Education. She has been awarded her Ph.D. degree in Computer Applications. She has completed her master's in computer applications (MCA) from Kurukshetra University. She is having 18 years of academic and research experience. She is a member of various professional bodies like ACM, IAENG and CSI. She has participated in various National and International Conferences and actively involved in various projects. There are more than 40 publications to her credit in reputed National and International Journals and Conferences. She is also member of Editorial board of various highly index journals. Her specialized domains include Mobile Computing, Web Applications, Data Mining and Software Engineering and guiding research scholars in these areas. She can be contacted at email: shaveta.fca@mriu.edu.in.



Mridula Batra    Faculty of Computer Applications, MRIIRS. She has participated in various National and International Conferences and actively involved in various projects. There are more than 20 publications to her credit in reputed National and International Journals and Conferences. She is also member of Editorial board of various highly index journals. She can be contacted at email: Mridula.fca@mriu.edu.in.