Development of framework for detecting smoking scene in video clips

Poonam G, Shashank B. N, Athri G Rao
Department of Computer Science and Engineering, Rashtreeya Vidyalaya College of Engineering (RVCE),
Bengaluru, India

Article Info

Article history:

Received Aug 21, 2018
Revised Oct 22, 2018
Accepted Nov 5, 2018

Keywords:

Deep learning
Faster R-CNN
Object detection
Smoking scene detection
Tensorflow

ABSTRACT

According to Global Adult Tobacco Survey 2016-17, 61.9% of people are quitting tobacco. The reason was the warnings displayed on the product covers, video clips, and advertisments. The focus of this paper is to automate the process of displaying warning messages in video clips. This paper explains the development of a system to automatically detect the smoking scenes using image recognition approach in video clips and then add the warning message to the viewer. The approach aims to detect the cigarette object using Tensorflow's object detection API. Tensorflow is an open source software library for machine learning provided by Google which is broadly used in the field image recognition. At present, Faster R-CNN (Region-based Convolutional Neural Networks) with Inception ResNet is theTensorflow's slowest but most accurate model. Faster R-CNN with Inception Resnet v2 model is used to detect smoking scenes by training the model with cigarette as an object.

Corresponding Author:

Poonam G,
Department of Computer Science and Engineering,
Rashtreeya Vidyalaya College of Engineering (RVCE), Bengaluru, India.
Email: poonamghuli@rvce.edu.in

1. INTRODUCTION

In this age of online and social media, cinemas remain a prominent form of entertainment and influence on youth. India produces approximately 800 to 1000 movies a year, leading to the requirement of displaying the warning message when the smoking scene is showcased. The warnings as of now are displayed manually.

This proposed work explains the development of a framework to automatically detect the smoking scenes using neural network model and then display the required warning message. The challenge to detect the smoking scenes in video clips is that only the small portion of the smoking event may be showcased and it may be displayed for fraction of a second. To overcome this challenge, object detection methods are used to detect different kind of cigarettes. These cigarettes may have varying shapes, colors and size. Then a warning message such as "Smoking Kills" or "Smoking is injurious to health" is displayed in the video clip.

Google's Object detection API is built on top of Tensorflow. There are different pre-trained Tensorflow models available for object detection such as Single Shot Multibox Detector (SSD) with MobileNet, SSD with Inception V2, Region-Based Fully Convolutional Networks (R-FCN) with Resnet 101, Faster R-CNN with Resnet 101 and Faster R-CNN with Inception Resnet v2. In our proposed approach Faster R-CNN with Inception Resnet v2 is chosen. Faster R-CNN has achieved much better speed and accuracy. Future models followed various approaches but could outperform Faster R-CNN by a significant margin. Faster R-CNN may not be the simplest or fastest method for object detection, but it is still one of the best performing. At present, Faster R-CNN with Inception ResNet model of Tensorflow is the slowest but most accurate model.

2. RELATED WORKS

Research work carried out in [1] exploits the Region Proposal Network (RPN) of the Faster R-CNN model to detect pedestrians. Even though R-CNN’s Region Proposal Network (RPN) performs well, the results can be degraded by downstream classifiers. The two main reasons that may lead to this situation are: handling of small instances due to the insufficient resolution of the features and the mining of hard negative cases is difficult due nearly no presenece if bootstraping methodologies to achieve the same.

Another research work carried out in [2] concentrates on the datasets's impact on deep learning and the application and the importance of deep learning through Faster R-CNNs. The work tries to summarize the deep learning algorithms and common data sets used in the field of computer vision. Additionally, the study builds a newer dataset in accordance to the previously available and commonly used datasets. Faster R-CNN is then applied over this newly built dataset.

Application of the faster R-CNN is explored on various benchmarks on which the algorithm has a proven improved results ranging from object detection to face recognition in [3]. The worked carried out provides the results of training a Faster R-CNN model on the large scale face dataset WIDER [4]. The work also tries to explain the results on WIDER dataset along with two more state of the art and widey used datasets FDDB and IJB-A.

Online handwritten graphics may contain mathematical expressions and diagrams. Detection of these symbols consist of methods designed for a single graphic type [5]. In this work, evaluation of the Faster R-CNN object detection algorithm as a general method for detection of symbols in handwritten graphics is carried out. Different configurations of the Faster R-CNN method are evaluated, and issues relative to the handwritten nature of the data are pointed out. Considering the online recognition context, evaluation of efficiency and accuracy trade-offs of using Deep Neural Networks of different complexities as feature extractors is carried out.

3. DEEP LEARNING MODELS FOR OBJECT DETECTION

Deep Learning is a part of machine learning which gives outstanding performance in the image and video classification tasks. In deep learning there are various architectures including recurrent neural networks and deep neural networks which have major applications in the field of computer vision, machine translation and natural language processing. Deep neural network architecture has multiple hidden layers between its input and output layers which are feed forward. Data from the input layer flows to the output layer without looping back. The main application of computer vision involves object detection which has a main focus on research. The progress in object detection is mainly because of Convolutional Neural Networks(CNN).

Object detection is improved from single object to multiple object detection in recent years. The first can detect a single object in an image which can be used for classification tasks. In the later approach not only multiple objects in an image are detected but also their exact location in the image is indicated by rectangular boxes or masks. Moving object detection has been explored extensively by various authors by the applying different methods [6-7] and so on. But most of the proposed works has considered stationary cameras or a stationary background. The CNN architecture is constantly improving from ALexNet [8], the ZF Net [9], the VGG Net [10], the ResNet [11] starting from the year 2012. An object detection algorithm was formed based on R-CNN in deep learning and a number of well-known datasets are considered to improve these algorithms with improvement in the accuracy of detection.

Various changes to the network structure has improved the deep learning that the network uses. The most well-known series of algorithms for object detection are based on R-CNN which include the following:

3.1 R-CNN (Region-Based Convolutional Neural Networks)

Images in the dataset are labelled or the region of our interest is cropped out from the image and this cropped region is given as input to the convolution neural network. When they are given as input to the network, application of rectangular bounding box regressor is provided. For the classification purpose SVM is used. In terms of both space and time, training becomes very expensive. The object detection in images is often slow and it takes around one minute per image.

3.2 Fast R-CNN

Fast R-CNN takes the characteristics of both R-CNN and SPP-Net. Fast R-CNN takes the entire image and feature map is created by forwarding it to the convolutional layer. Then region of interest is found by RoI pooling layer. RoI layer is a single SPP layer that is applied on top of convolutional layer which is then attached to fully connected layer. Bounding box regressors and softmax classifiers are applied for classification. Based on both bounding box regressors and softmax classifiers, multitask loss is computed. By this, the layer below single SPP layer is made trainable and the problem associated with SPP-Net is solved. It higher detection quality in the main improvement done over R-CNN and SPP-Net. Here all the layers can be updated during training process and it does not require the features to be stored in a disk. The Fast R-CNN training time is 9 times faster when compared to R-CNN which is 3 times faster than SPP-Net and testing time required is 213 times faster than R-CNN and when compared to SPP-Net it s 10 times faster. Along with the decrease in training time, there is increase in the level of accuracy.

3.3 Faster R-CNN

Regional Proposal Network takes image of any size as input and output a set of object proposals each with objectness score [12-13]. SPP-Net and Fast R-CNN has reduced the execution time of object detection but more time is required for regional proposal. Faster R-CNN solves this problem by using deep networks for traditional practices to compute a proposal box. Faster R-CNN consists of two modules. First is the fully connected convolutional network and the second is Fast R-CNN detector.

4. PROPOSED METHODOLOGY

The following section gives the details of the methodology used in the proposed work.

4.1 Experimental Setup

The details of hardware chosen for our experiment are shown in Table 1. Experimental setup requires for which a system with minimum 2 GB of NVIDIA GPU card, CUDA and cuDNN installed. The model uses the Anaconda Python and the python packages including Tensorflow, OpenCv, matplotlib and pandas.

Table 1. Hardware Requirements

Hardware components	Configuration
Processor	Intel i5 7200
Processor Speed	2.6 GHz
RAM size	8 GB
OS	Windows 10
GPU	NVIDIA GeForce GTX 940M
VRAM	2 GB

4.2 Dataset

Hundreds of images are required to train the classifier for good detection. Videos containing smoking scenes are collected. These videos contain cigarettes with different shape, size and color. From these videos 5 frames are extracted per second. All these frames are converted to 200 X 200 JPEG images which forms our training dataset. The training dataset contains 660 images. The cigarettes in the images have variety of lighting conditions and backgrounds. Also there are images in which cigarette is partially seen. Using Labeling tool, the location of cigarette in an image is marked by drawing rectangles. The location of cigarette is stored in an XML file which contains information about image height, width and the coordinates of the bounding rectangle drawn. Each image size is less than 100KB since the time required for training becomes large if the image size is high.

4.3 Training

Training the model for detecting the cigarettes can be done on Google cloud services, CPU or GPU. During training, for a particular time interval Tensorflow stores the checkpoints. Loss is reported in each step of training. The loss reported by the model is a combination of classification loss and regression loss. Training the classifier is stopped when loss is dropped to 0.04. The latest checkpoint created by Tensorflow at a loss of 0.4 is used for detection the cigarettes.

5. RESULTS

The model can input in 3 different forms: image, video or a live webcam feed. The system results are evaluated using images and videos. First the system was evaluated using different datasets with different images. Dataset 1 contain 10 images among which 5 images have cigarettes. Dataset 2 contain 20 images among which 10 images have cigarettes. Dataset 3 contain 30 images among which 15 images have cigarettes. Accuracy, sensitivity and specificity are the performance measures considered to evaluate the model with these datasets. The results are shown in Table 2. Since datasets 2 and 3 contain images where cigarette is partially visible and also due to illumination changes in the image, the accuracy is reduced.

Table 2. Evaluation Results

Datasets	Performance measures		
	Accuracy	Sensitivity	Specificity
Dataset 1	90%	80%	100%
Dataset 2	75%	70%	80%
Dataset 3	76%	73%	80%

Next the model is evaluated using a video dataset that contain 10 videos. The results are displayed in Table 3. The proposed approach gives an average accuracy 94.08 for the video dataset considered.

Table 3. Result Analysis of Videos

Sl. No.	Details of confusion matrix					
	Frame size	True posttives	True negatives	False positives	False negatives	Accuracy (%)
Video 1	244	210	24	6	4	95.90
Video 2	92	83	6	2	1	96.74
Video 3	303	276	16	3	8	96.37
Video 4	483	417	36	7	23	93.79
Video 5	351	278	39	12	22	90.31
Video 6	132	119	9	0	4	96.97
Video 7	567	401	77	19	10	84.30
Video 8	821	733	81	0	7	99.14
Video 9	573	423	99	8	43	91.10
Video10	885	763	88	9	25	96.16

Our results are compared with the smoking event detection ratio histogram method [13]. The comparision is shown in Table 4.

Table 4. Comparision of Results

Dataset	Details of confusion matrix				
	Frame size	True posttives	False positives	False negatives	Accuracy (%)
Videos considered in histogram method	2196	1824	30	120	93.2
Our video dataset	4451	4003	66	207	93.87

The results prove that object detection through faster R-CNN can be used for detection of smoking scenes by considering cigarette as an object.

6. CONCLUSION

The results show that proposed method can be adopted for displaying warning messages during smoking scenes. Our proposed work displays warning message by detecting cigarettes. This work can be extended to detect smoking scenes which do not contain a cigarette but exhaling the smoke. In Indian movies and television shows, a similar kind of messages are displayed during the event of alcohol consumption for this proposed work can be extended.

REFERENCES

[1] Zhang, Liliang, et al. “Is faster R-CNN doing well for pedestrian detection?”. *European Conference on Computer Vision. Springer*, Cham, 2016; 443-457.

[2] Zhou, Xinyi, Wei Gong, WenLong Fu, Fengtong Du. “Application of deep learning in object detection”. *Computer and Information Science (ICIS)*. IEEE/ACIS 16th International Conference on, IEEE. 2017; 631-634.

[3] H. Jiang and E. Learned-Miller. “Face Detection with the Faster R-CNN”. 2017 *12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. Washington, DC, 2017; 650-657.

[4] S. Yang, P. Luo, C. C. Loy, and X. Tang. "WIDER FACE: A face detection benchmark", *CVPR,* 2016.

[5] F. D. Julca-Aguilar and N. S. T. Hirata, "Symbol Detection in Online Handwritten Graphics Using Faster R-CNN", 2018 *13th IAPR International Workshop on Document Analysis Systems (DAS)*, Vienna, Austria, 2018, 151-156.

[6] Xu, P. "Study on Moving Objects by Video Monitoring System of Recognition and Tracing Scheme". *Indonesian Journal of Electrical Engineering and Computer Science(IJEECS)*. 2013; 11(9), 4847-4854.

[7] Mengxin Li, Jingjing Fan, Ying Zhang, Rui Zhang, Weijing Xu, Dingding Hou1. "Moving Object Detection and Tracking Algorithm". *Indonesian Journal of Electrical Engineering and Computer Science(IJEECS)*. 2013; 11(10), 5539 – 5544.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton. "ImageNet classification with deep convolutional neural networks". *NIPS*,2012.

[9] M. D. Zeiler and R. Fergus. "Visualizing and understanding convolutional neural networks". In *ECCV,*

[10] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". *ICLR*. 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". *CVPR,* 2016.

[12] Ren, Shaoqing, Kaiming He, Ross Girshick, Jian Sun. "Faster R-CNN: towards real-time object detection with region proposal networks". *IEEE transactions on pattern analysis and machine intelligence*. 2017; 39 (6), 1137-1149.

[13] Wu, Pin, Jun-Wei Hsieh, Jiun-Cheng Cheng, Shyi-Chyi Cheng, Shau-Yin Tseng. "Human smoking event detection using visual interaction clues". *20th International Conference on Pattern Recognition (ICPR) IEEE*, 2010; 4344-4347.