International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos

***[1]D Navaneetha,[2] K Gangamani, [3]A Hymavarshini**

[1]Associate Professor, Department of Information Technology, Bhoj Reddy Engineering College for Women, Hyderabad, India

[2,3]Students, Department of Information Technology, Bhoj Reddy Engineering College for Women, Hyderabad, India

ARTICLEINFO

Article History:

Accepted: 10 April 2023
Published: 30 April 2023

Publication Issue
Volume 9, Issue 2
March-April-2023

Page Number
666-671

ABSTRACT

With the emergence of screened films, Video content classification has become ubiquitous. Television films and Internet sites films are a big source of violence that may psychologically hurt teenagers. Although recently, Deep learning video classification has been developed quickly, a Comprehensive survey is needed to summarize the previous work done in this field. Therefore, this survey paper shows the common methods used in video classification. We further discuss the importance of filtering sensitive content such as (pornography, violence, gory, etc.) because of the increasing consumption of films by people of all ages. Several real-world verdict cases are similar scenarios to films with many scenes of violence. As deep learning has shown big success in computer vision areas, researchers are giving it a lot of attention.

Keywords: Video filtering; Video analysis; Video classification

I. INTRODUCTION

Videos can be categorized into many categories. It is possible to classify them according to the content, such as educational or entertainment content, or to classify them according to the type of images from which these videos were prepared, such as two-dimensional films, three-dimensional films, and cartoon films. As for the video content, there is much content that should be filtered to protect children, including sexual content, violence and the content that causes autism. We will present some previous attempts aimed at filtering some or all of the inappropriate content from some types of videos. Pornography has many common definitions in the psychological and scientific fields. One of the most known definitions states that pornography [1] is any commercial product in the form of fictional drama designed to elicit or enhance sexual arousal. Another definition states that it is any visual or printed content that contains the display of sexual activities or organs or explicit description, intended to stimulate sexual excitement and whatever you choose from the previous definitions, they all agreed that pornography is all about sexual excitement and show the bounds. Violence detection is a strenuous problem due to the heterogeneous content and variable quality of videos. Supervised classification is a fundamental task in machine learning. violent scenes are associated with

nude colour in video frames and groans and moans in the audio. there are cases like wrestling in which we may have false-negative or false-positive results. Traditional video filtering methods only work on a single dimension of features such as video frames colour analysis or video. When multiple dimensions of features such as (image frames colours, audio content, motion in the frame sequence of video, or emotions of the audience) are used, how can these features be integrated to perform accurate classification? The existence of such features raised the need of multi-feature learning [2], [3] [4]. Colour red for multi-feature classification, it may be required to identify classes of subjects that differ in each of the data views. In the past decades, video content filtering has attracted more and more attention, so it is necessary to summarize the state of the art and outline open problems and future enhancements. We divide the video filtering methods into model-based approaches and similarity-based approaches. Generative approaches learn the distribution of the features and use generative models to represent video classification. Discriminative approaches optimize a function that tries to keep down the different classes' average similarity. Discriminative approaches have many types based on the combination method of the multi-feature information such as common eigenvector matrix or common indicator matrix. we can also divide previous work done into two categories one of them is the targeted content to be filtered such as (violence only, pornography only or both of them), the second category can be the targeted media type such as (real videos only, animation cartoons only or both of them), as shown in fig [1].

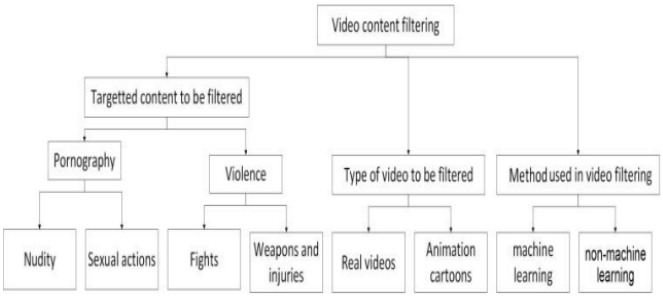

Fig.1. Types of Inappropriate content filtering

As far as what this paper is concerned, the video filtering and video analysis papers of are published in top machine learning venues like the International Conference on Machine Learning (ICML) [5], Neural Information Processing Systems (NIPS) [6], IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) [7], International Conference on Computer Vision (ICCV) [8], Association for the Advancement of Artificial Intelligence (AAAI) [9], International Joint Conference on Artificial Intelligence (IJCAI) [10]. Although video filtering has shown considerable success in practice, some open problems limit its advancement. We explain several open problems and hope that readers can have a better version of the automatic video filtering using deep learning.

II. RELATED WORK

Among all media types (e.g., texts and audio), images are the most used pornography carrier. Since pornography often has skin exposure, skin detection is commonly used for pornography detection. However, in order that skin detection is challenging, the Approaches used have a limited ability of generalization. Vu Lam and Duy-Dinh Le introduced MediaEval that combined the trajectory-based motion features with SIFT-based and audio features. their results show that the trajectory-based motion features still have very competitive performance. the combination with image features and audio features can improve overall performance for detection of violence scenes in videos. Paper states that for video action recognition recently dense we can use trajectories and it also can be used for video representation. they use the camera motion to improve performance. they use SURF descriptors to estimate motion of camera and match between frames and feature points. Unlike camera motion, Human motion generates inconsistent matches. A Human detector is used to increase accuracy of the estimation and to remove any inconsistent matches. Given the estimated

camera motion, paper removes trajectories consistent with it and uses this estimation to cancel out camera motion from the optical flow. This significantly improves motion-based descriptors, such as HOF and MBH. Focusing on filtering real-time videos, Nevenka and Radu produced an effective patent that automatically filters multimedia program content in real-time based on stock and user-specified criteria. Multiple multimedia processors are used to analyze audio or visual content. At the end, It should analysis the resultant and compare it to a specific selected criteria.

In paper, nude detection is implemented with the use of two algorithms. The first algorithm detects humans from the processed frames and then crop them out. Then the second algorithm is the nude detection algorithm used to determine if the images are nudes or not. the technique used in this paper has many problems. first is that while detecting nudity, it fails to properly classify black and white images as containing nudity or not because this paper is based mainly on skin colour. When a person is wearing a skin-toned dress the paper's algorithm gives a false positive result and it also fails to detect nudity if most of the body portion is not included in the video frames. paper presents a system developed for contentbased news video browsing for home users. This system integrates the audio-visual as well as text detection and NLP technique analysis to extract structure and content information of news video and to organize and categorize news stories. Jin used weighted multiple instances learning train a generic region-based recognition model on images modelled as a bag of the region, taking into account the degree of pornography for each region. He shows that the key pornographic contents often are located in local regions in an image, and the background regions may be destructive. He stated that there are two main pornographic contents: private body parts, and sexual behaviours. We perform bounding box annotation for these key pornographic contents in the training set, and for the sexual behaviour, using annotations, there are only a small number of annotations (less than 10) required for both types of pornographic contents. each image X is modeled as a bag of n regions xi|i = 1..., n. Given a region, the deep CNN extracts layer-wise representations from the first layer which is convolutional to the fully connected layer as the last one. His CNN architecture is inspired by the GoogLeNet model.

III. PROPOSED SYSTEM

The challenge for human censors to keep up with the vast volume of user-generated material on the site is the issue that the suggested approach for detecting improper content in YouTube videos using deep learning seeks to overcome. The enormous volume of movies being submitted every minute cannot be handled by the conventional techniques of content management, such as manual review and keyword identification. Additionally, producers who utilize euphemisms or misspellings to evade detection can easily get around keyword-based detection systems. For YouTube moderators, who are responsible with screening video and flagging offensive material before it is made public, this presents a big difficulty. In order to increase the effectiveness and precision of content moderation on the platform, the suggested solution automates the process of finding offensive content in YouTube videos using deep learning. This might make the internet a safer place for all users, especially kids who could be more susceptible to unsuitable information.

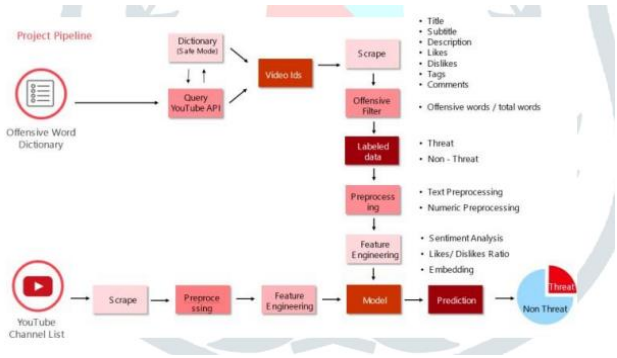

Fig 2: System Architecture

TRAINING METHODOLOGIES

Several training methodologies are used in the proposed methodology for utilizing deep learning to find objectionable content in YouTube videos. The following are some of the main methods that might be applied:

A. Data Collection: Collecting a meaningful and varied dataset is the initial stage in every machine learning effort. It would be necessary to gather a sizable dataset of YouTube videos for this strategy, along with proper labels for the many categories of objectionable content that would be recognized (e.g., nudity, hate speech, violence, etc.). The dataset should be balanced in terms of the proportion of both good and bad instances, and it should be representative of the kinds of content that are often submitted to YouTube.

B. Data Preprocessing: The gathered dataset would undergo preprocessing to remove the pertinent data. This might entail both extracting audio features from the movies, such as spectrograms or MFCCs, as well as visual information from the films, such as color histograms, texture features, or motion vectors. The preparation procedure may also include data cleansing, the elimination of superfluous data, and feature normalization. We have collected and preprocessed 37,354 YouTube videos based on the content, visual and audio.

C. Model Selection: A suitable deep learning model would be chosen for the classification job after the dataset has been preprocessed. Choosing from a variety of pre-trained models, such as CNNs, RNNs, or GANs, or creating a unique architecture that is suited to the particular task, might be involved. In addition to its performance on comparable tasks, the model should be chosen for its computational effectiveness and scalability.

D. Model Training: The chosen model must next be trained using the preprocessed data. In order to do this, the dataset must be divided into training, validation, and test sets. The model must then be trained using the training set and a suitable loss function and optimization technique. To track the model's performance and avoid overfitting, periodic validation set evaluations should be performed. To enhance the model's performance, hyperparameter adjustment may also be used.

E. Model Evaluation: To assess the model's performance on unobserved data, a test set evaluation must be conducted once the model has been trained. A qualitative review of the kinds of errors the model produced as well as metrics like precision, recall, F1-score, and accuracy might be included in the evaluation. To assess the model's efficacy, it might be contrasted with other cutting-edge techniques for finding objectionable material in films.

F. Deployment: After being trained and assessed, the model might be used to instantly detect objectionable material in YouTube videos. The model might be used to automatically flag or delete videos that contain objectionable content and be integrated into the YouTube or other video-sharing platform's content moderation system.

IV. CONCLUSION AND FUTURE WORK

This research proposes a revolutionary deep learning-based system for the identification and categorization of unsuitable video material. To extract the characteristics of movies, transfer learning utilizing EfficientNet-B7 architecture is used. The BiLSTM network processes the retrieved video features and conducts multiclass video classification while the model learns the efficient video representations. A dataset of 37,753 carefully annotated cartoon video clips that were downloaded from YouTube is used for all assessment studies. The evaluation results showed that compared to other tested models, such as Efficient Net-FC, Efficient Net-SVM, Logistic Regression, Decision tree, and Efficient Net-BiLSTM with attention mechanism-based models, the proposed framework of Efficient Net-BiLSTM (with hidden units D 128) exhibits higher performance (accuracy 96%) (with hidden units D 64, 128, 256, and 512). Furthermore, by earning the highest recall score of

92.22% in the performance comparison with current state-of-the-art models, our BiLSTM-based framework outperformed other existing models and approaches. The following are some benefits of the proposed deep learning-based approach for identifying unsuitable video content for children:

 It filters live-captured videos by analysing the footage at a speed of 22 frames per second while taking into account the current conditions. Its deep learning framework is built on EffcientNet-B7 and BiLSTM.

 It can help any video-sharing site either eliminate any clips that are hazardous or blur/hide any frames that are uncomfortable.

 It may also aid in the creation of browser add-ons or plugins that automatically filter out information that is inappropriate for children in order to provide parental control solutions for the Internet.

Furthermore, our method for locating inappropriate children's content on YouTube does not rely on the metadata of the video, which may be easily altered by malicious uploaders in an effort to deceive users. We suggest merging the temporal stream using optical low frames with the spatial stream of the RGB frames in order to improve the model's performance by understanding the overall representations of movies. We also wish to enhance the classification labels in order to better target the many types of inappropriate children's content seen in YouTube videos.

V. REFERENCES

[1]. L. Ceci. YouTube Usage Penetration in the United States 2020, by Age Group. Accessed: Nov. 1, 2021. [Online]. Available:https://www.statista.com/statistics/296227/us-youtube-reach-age-gender/

[2]. P. Covington, J. Adams, and E. Sargin, ``deep neural networks for YouTube recommendations,'' in Proc. 10th ACM Conf. Recommender Syst., Sep. 2016, pp. 191_198, doi: 10.1145/2959100.2959190.

[3]. M. M. Neumann and C. Herodotou, ``Evaluating YouTube videos for young children,'' Educ. Inf. Technol., vol. 25, no. 5, pp. 4459_4475, Sep. 2020, doi: 10.1007/s10639-020-10183-7.

[4]. J. Marsh, L. Law, J. Lahmar, D. Yamada-Rice, B. Parry, and F. Scott, Social Media, Television and Children. Shef_eld, U.K.: Univ. Shef_eld, 2019. [Online]. Available: https://www.stac-study.org/downloads/ STAC_Full_Report.pdf

[5]. L. Ceci. YouTube_Statistics & Facts. Accessed: Sep. 01, 2021. [Online]. Available: https://www.statista.com/topics/2019/youtube/

[6]. M. M. Neumann and C. Herodotou, ``Young children and YouTube: A global phenomenon,'' Childhood Educ., vol. 96, no. 4, pp. 72_77, Jul. 2020, doi: 10.1080/00094056.2020.1796459.

[7]. Z. Guo, L. Tang, T. Guo, K. Yu, M. Alazab, and A. Shalaginov, “Deep Graph neural network-based spammer detection under the perspective of heterogeneous cyberspace,” Futur. Gener. Comput. Syst., vol. 117, pp. 205–218, 2021, doi: 10.1016/j.future.2020.11.028.

[8]. D. C. Corrales, A. Ledezma, and J. C. Corrales, “A case-based reasoning system for recommendation of data cleaning algorithms in classification and regression tasks,” Appl. Soft Comput. J., vol. 90, p. 106180, 2020, doi: 10.1016/j.asoc.2020.106180.

[9]. A. Fahfouh, J. Riffi, M. Adnane Mahraz, A. Yahyaouy, and H. Tairi, “PV-DAE: A hybrid model for deceptive opinion spam based on neural network architectures,” Expert Syst. Appl., vol. 157, p. 113517, 2020, doi: 10.1016/j.eswa.2020.113517.

[10].S. Panda, A. K. Ghosh, A. Das, U. Dey, and S. Gupta, “Machine Learning-based Linear regression way to deal with making data science model for checking the sufficiency of night curfew in Maharashtra , India,” vol. 1, no. 2, pp. 168–173, 2021.

[11].A. Kantchelian, J. Ma, and A. D. Joseph, “Robust Detection of Comment Spam Using Entropy Rate Categories and Subject Descriptors,” no. AISec, pp. 59–69, 2012.

[12].E. Tan, L. Guo, S. Chen, X. Zhang, and Y. E. Zhao, "Spammer Behavior Analysis and Detection in User Generated Content on Social Networks," 2012, doi: 10.1109/ICDCS.2012.40.

Cite this article as :

D Navaneetha, K Gangamani, A Hymavarshini, "A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 2, pp.666-671, March-April-2023.