International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 11 ǁ November 2023 ǁ PP. 01-05

Inappropriate Content and Misleading Thumbnail Detection of YouTube Videos via Deep Learning

Sneha Giridharan

*[*1]Department of Computer Science and Engineering, Govt Model Engineering College , Kochi*

Abstract

The YouTube platform has experienced tremendous growth in all dimensions of expansion. Every minute, 500 hours of videos are uploaded to YouTube, attracting billions of active viewers and transferring petabytes of data. YouTube also consists of children and youngsters under the age of 18. Moreover, Kids also contribute a big part to the viewers of all videos on YouTube. Being protected by the COPA Act in the US, kids are still getting good exposure to video materials intended for adults. The minds of kids and children are finding ways to circumvent the protections kept by YouTube. Due to the vulnerable mind of children, the content can cause shock, acute trauma, and even PTSD.

The framework finds mature and gore content from a YouTube video and parses it through a pre-trained convolutional neural network [CNN]. The frames of the video are then sent to Bi-directional Long Short Term Memory Network to find mature and gore content and flag them as Not Safe for Children. The Thumbnail of a video also plays a vital role in tricking kids into watching the Video, generally known as click-baiting. Thumbnails are downloaded using YouTube API and then processed through a pre-trained CNN model, which predicts if it matches with any of the frames in the Video; if it is not present in the Video, The framework flags the Thumbnail as Click-bait.

Keywords: *Video Classification, Convolutional Neural Networks, Deep Learning, Social Media Analysis*

Date of Submission: 20-10-2023 Date of acceptance: 03-11-2023

I. INTRODUCTION

The YouTube platform has experienced tremendous growth in all dimensions of expansion.Every minute, 500 hours of Video are uploaded to YouTube, attracting billions of active viewers. YouTube viewers consists of children and youngsters under the age of 18.Kids also contribute a big part to the viewers of all videos on YouTube. Being protected by the COPA Act in the US, kids are still getting good exposure to video materials intended for adults. Due to the vulnerable mind of children, the content can cause shock, acute trauma, and even PTSD[1].

The framework finds mature and gore content from a YouTube video and parses it through a pre-trained convolutional neural network [CNN].The frames of the Video are then sent to Bi-directional Long Short Term Memory Network to find mature and gore content and flag them as Not Safe for Children.The Thumbnail of a video also plays a vital role in tricking kids into watching the Video, generally known as click-baiting. Thumbnails are downloaded using YouTube API and then processed through a pre-trained CNN model.Then predicts if it matches with any of the frames in the Video.if it is not present in the Video, The framework flags the Thumbnail as Clickbait.

1.1.1 System Study Report

A deep model based approach is explores for mature content and click-bait detection. Specifically, a system is designed and implemented for inappropriate content as well as click-bait detection that produces the result from YouTube Videos. The advantages and disadvantages of a number of previous works have been studied and an advanced automatic face recognition system is proposed to identify faces from images.

1.1.2 Motivation

The exponential growth of videos on YouTube has attracted large community of viewers especially children. Large number of up-loader's find this as one of the best opportunity to spread toxic visual content with children. Therefore we need a system to overcome this toxicity.

[2]There are large number of filters available that can be used to filter out the mature content in the YouTube data. But most of them are not that efficient in filtering out all the gore content. These content can easily escape from the available filters and can cause problems. Therefore we need a filtering system that is enough to overcome this toxicity.

1.2 PROPOSED SOLUTION

The product identifies the videos which are not intended for children, Videos which contain mature and violent contents can cause adverse effects on children's minds. This can lead to suicidal thoughts, low self-esteem, aggressive behaviour.

The application will be operating in a containerized, cloud environment with access to Graphics Processing Units, in a Linux environment, No data will be retained in the servers after operation. The Application will have a web interface, where users will be interacting and Web Application Servers and Machine Learning Model in the server.The application is designed with security and privacy concerns in mind. The Application does not retain any content created by the content developer in the server. Only the log is kept back in the server. Few constraints that could arise is memory limitation of cloud based approach. Cloud Native Computing could be a financial constraint.

1.2.1 CNN Architecture

The architecture used in this code is a transfer learning architecture that leverages a pre-trained ResNet50V2 model as a feature extractor and adds a new output layer for binary classification. This type of architecture is commonly referred to as a fine-tuned pre-trained model or a transfer learning model.\\

The model architecture is based on the pre-trained ResNet50V2, which is a deep convolutional neural network with 50 layers. The ResNet50V2 has an input tensor with the shape of (128, 128, 3). On top of the ResNet50V2 base model, a Dense layer with one neuron and a sigmoid activation function is added. This layer serves as the output layer for the binary classification task.

The model has a total of two trainable layers: the added Dense layer and the last layer of the ResNet50V2, which is also a Dense layer. All the other layers in the ResNet50V2 are frozen and not trainable. The loss function used for this model is binary cross-entropy, and the optimizer used is Adam.

1.2.2 Layer Wise Structure

In a neural network, there are two main types of layers: trainable layers and frozen layers. Trainable layers are those whose weights can be updated during training, while frozen layers have fixed weights that are not updated during training.

In this specific model, the ResNet50V2 network is used as a pre-trained feature extractor. This means that the ResNet50V2 has already been trained on a large dataset (ImageNet) to identify and extract features from images, such as edges, shapes, and textures. Therefore, to avoid overfitting and preserve the learned features, the weights of the ResNet50V2's layers are frozen, and only the weights of the last layer of the ResNet50V2 and the added Dense layer are updated during training.

The added Dense layer is used as the output layer for the binary classification task. The Dense layer has one neuron and a sigmoid activation function. The sigmoid function takes in a value and squashes it between 0 and 1, which is useful for binary classification because it can be interpreted as the probability that an image belongs to one of the two classes. The weights of this Dense layer are initialized randomly and are updated during training to optimize the model's performance.

The last layer of the ResNet50V2 is also a Dense layer, and its weights are fine-tuned during training to adapt to the specific binary classification task. By fine-tuning the weights of this layer, the model can learn to recognize the specific patterns and features that are important for the binary classification task at hand.

The binary cross-entropy loss function is used to measure how well the model is doing on the binary classification task. The goal during training is to minimize the loss function by adjusting the weights of the trainable layers in the model. The optimizer used to adjust the weights is Adam, which is a popular optimization algorithm that works well for deep learning models. Overall, the combination of frozen and trainable layers, along with the binary cross-entropy loss function and the Adam optimizer, allows the model to efficiently learn and classify images into one of two classes.

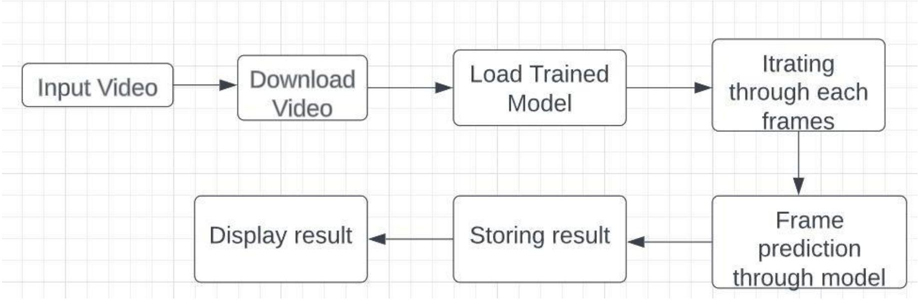
1.2.3 Component Level Architecture

i Clickbait Detection Engineii Mature Content Detection

Mature Content Detection

One way to detect mature content in YouTube videos is by using an artificial intelligence (AI) model.we need to design features and train a model to predict whether a given video contains mature content.Some possible features include certain words or phrases in the video's title or description, the use of a particular image or video content, or certain metadata tags.To train the classifier, you would need a labeled dataset of YouTube videos that have been manually labeled as either containing mature content or not containing mature content. Once the classifier is trained,it is used to predict whether a given YouTube video contains mature content or not.

Figure 1: Mature Content Detection

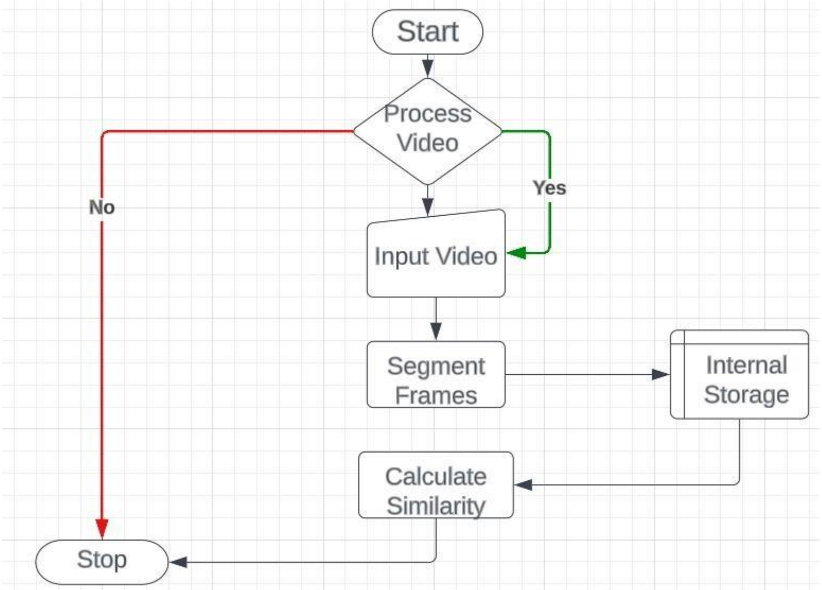


Click-bait Detection Engine

Thumbnail images play an essential role in attracting viewers to a video.They provide a quick visual summary of the content and can help to grab the attention of users who are scrolling through search results or browsing a channel's page. Click-bait thumbnails are thumbnail images that are designed to entice viewers to click on a video by using sensational or misleading images or headlines. Click-bait thumbnails can be annoying and frustrating for users who feel like they have been tricked into clicking on a video. Click-bait thumbnails are often used to increase the number of views on a video, as more views can lead to more ad revenue for the video's creator.

Artificial intelligence (AI) model is used to identify clickbait thumbnails.A pre-trained ResNet50V2 model is used to identification. Each image frames of the video are fed to the network to detect similarity.The percentage of similarity between the frames and thumbail is calculated.If the similarity exists beyond the threshold, then it is not a clickbait.

Figure 2: Click-bait Detection Engine



II. RESULT AND DISCUSSION

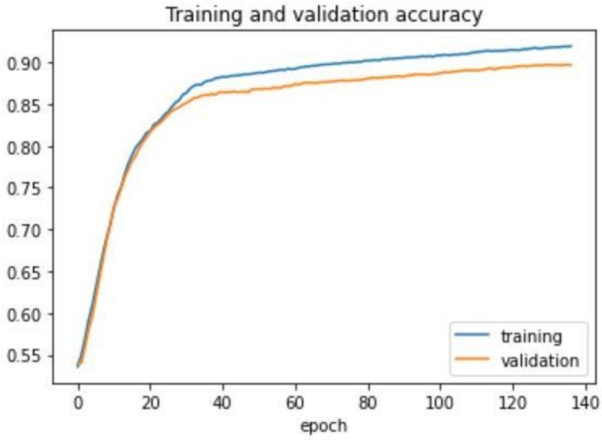
The results obtained are as discussed below

1.3.1 Dataset

For the experiment, 1000 violent and 1000 non violent videos are collected. The videos are split into frames for training as well as testing. Among the frames 70 percent of the frames are used for the training and remaining 30percent for the testing.

The best result in 137th epoch. The model scored about 90 percent accuracy.
The training accuracy is about 0.9196265339851379
The testing accuracy is about 0.8988044261932373.
Loss on training is about 0.21429312229156494
Loss on testing is about 0.25044387578964233.

Figure 3:Training and Validation Accuracy



While considering the model evaluation The correct predictions is about 2103 and the wrong predictions is about 239. The confusion matrix of the given model evaluation is given below. From the matrix, we can conclude that the system got about 90 percentage of accuracy.

Figure 4:confusion Matrix

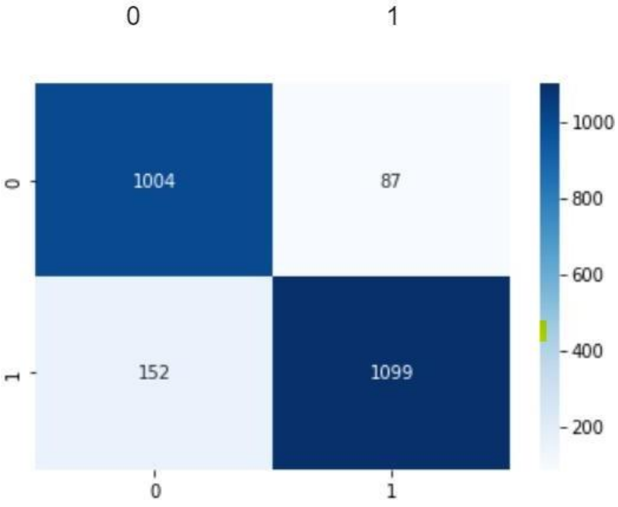
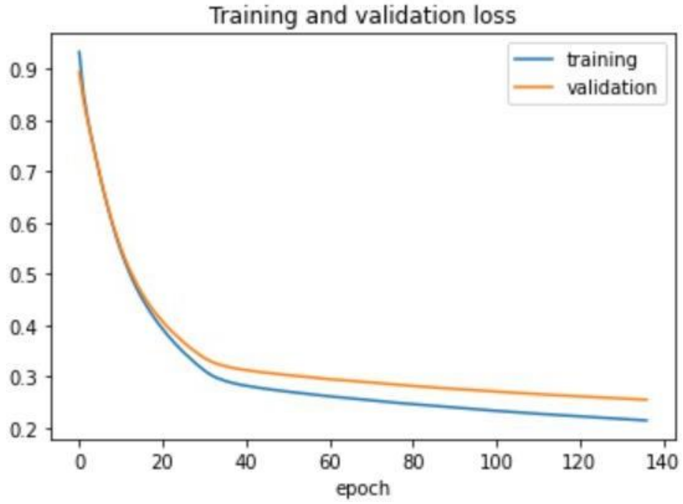


Figure 5: Loss Graph



III. CONCLUSION

Designed a model that can be utilized to detect a mature content and click-bait in YouTube Videos. The whole project can be expanded with more datasets and can be used for detecting mature content and click-bait from YouTube shorts.

REFERENCES

[1]. Paul Covington, Jay Adams, Emre Sargin "Deep Neural Networks for YouTube Recommendations" 2September 15-19, 2016,doi: http://dx.doi.org/10.1145/2959100.2959190

[2]. Sulthan Ashramani, Detecting and Measuring the Exposure of Children and Adolescents to Inappropriate Comments in YouTube, https://doi.org/10.1145/3340531.3418511

[3]. Rishabh Kaushal, Srishty Saha, Payal Bajaj, Ponnurangam Kumaraguru,KidsTube: Detection, Characterization and Analysis of Child Unsafe Content & Promoters on YouTube,https://doi.org/10.48550/arXiv.1608.05966

[4]. ZW. Han and M. Ansingkar, "Discovery of Elsagate: Detection of Sparse Inappropriate Content from Kids Videos," 2020 Zooming Innovation in Consumer Technologies Conference (ZINC), 2020, pp. 46-47, doi: 10.1109/ZINC50678.2020.9161808.

[5]. Rashid Tahir,Faizan Ahmed,Hammas Saeed,Shiza Ali,Fareed Zaffar, Christo Wilson,"Bringing the kid back into YouTube kids: detecting inappropriate content on video streaming platforms",doi:https://doi.org/10.1145/3341161.3342913

[6]. Detection of Sparse Inappropriate Content from Kids Videos," 2020 Zooming Innovation in Consumer Technologies Conference (ZINC), 2020, pp. 46-47, doi: 10.1109/ZINC50678.2020.9161808.

[7]. N. Aggarwal, S. Agrawal and A. Sureka, "Mining YouTube metadata for detecting privacy invading harassment and misdemeanor videos," 2014 Twelfth Annual International Conference on Privacy, Security and Trust, 2014, pp. 84-93, doi: 10.1109/PST.2014.6890927.

[8]. Alshamrani, Sultan, et al. "Investigating Online Toxicity in Users Interactions with the Mainstream Media Channels on YouTube." CIKM (Workshops). 2020.

[9]. Varshney, D., Vishwakarma, D.K. A unified approach for detection of Clickbait videos on YouTube using cognitive evidences. Appl Intell 51, 4214–4235 (2021). doi: doi.org/10.1007/s10489-020-02057-9

[10]. Kanwal Yousaf and Tabassam Nawaz,"A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos,vol.10, doi: 10.1109/ACCESS.2022.3147519