Detecting Inappropriate Videos for Children Using Deep Learning

Name1, Name2

***Abstract*—With the advent of child-centric content-sharing platforms, such as YouTube Kids, millions of children, from all age groups are consuming gigabytes of content, if not more, on a daily basis. With PBS Kids, Disney Jr. and countless others joining in the fray, this consumption of video data stands to grow further in quantity and diversity. However, it has been observed increasingly that content unsuitable for children often slips through the cracks and lands on such platforms. To investigate this phenomenon in more detail, we collect a first of its kind dataset for inappropriate kids videos hosted on such children-focused apps and platforms. Alarmingly, our study finds that there is a noticeable percentage of such videos currently being watched by kids with some inappropriate videos having thousands of views already. To address this problem, we develop a deep learning system that can flag such videos and report them. Our results show that our techniques can be successfully applied to various types of animations, cartoons and CGI videos to detect any inappropriate content within them.**

I. Introduction

YouTube is by far the most widely used video sharing platform today. People from all age groups, spend a noticeable amount of time each week browsing through the vast collection of videos available [1] [2]. This enormous and diverse audience of YouTube constitutes a significant number of children as well. In fact, it has become so ubiquitous amongst kids that according to a recent study [3], kids under 8 years old spend 65% of their time online watching YouTube videos. Hence, to cater to the needs of this ever expanding base of under aged viewers and address the concerns of their parents/guardians, YouTube developed a separate video sharing platform dedicated to children known as YouTube Kids (YTK). YTK, and other similar platforms and apps, such as Nick Jr., Disney Jr., PBS Kids etc., only contain content deemed appropriate for children under a certain age group. As a result, parents feel increasingly more confident giving their kids independence on what they want to view often under no supervision.

However, as this study highlights and numerous other reports confirm [4], [5], [6], [7], inappropriate content often lands on YouTube Kids. From videos containing references to sex and alcohol to drug use and pedophilia, a wide range of unsuitable content has been reported on YTK. In fact, some children-centric organizations and watchdogs have reported the YTK app to the Federal Trade Commission (FTC) in the past, alleging that they are deceiving parents by marketing YTK as being safe for children when in fact it is not [8], [9]. Of course, the intent behind the app is noble and YTK has deployed numerous automatic checks, such as filtering algorithms, and manual review mechanisms to filter out unwanted content. However, we assert that improvements can be made leveraging deep learning methods that can augment the filtering mechanisms deployed by Google and further reduce the amount of inappropriate content that manages to slip through the cracks.

To this end, we first need to develop a firm understanding of the depth and scope of problem. For simplicity, we broadly classify unsuitable content into two categories; Fake and Inappropriate. Fake videos are those where an uploader leverages a popular cartoon character or movie, such as Mickey Mouse, and creates additional content for their purposes. This fake video in of itself is not a problem for parents, as the video could be entirely harmless and show Mickey Mouse playing quidditch with Harry Potter. However, should the creator of the video desire they can very well leverage the popularity of the cartoon character and misuse it for sending the wrong messages. This could include Mickey Mouse advertising a particular product or making sexual advances towards Minnie Mouse. The latter would also fall into the category of an inappropriate video, which is defined as a video having explicit content deemed unsafe for children, examples of which include violence, nudity, gore etc. For our purposes, we have limited our study of inappropriate videos to those containing sexually explicit themes.

The issue of adult content on YTK has recently gained more traction. For instance, Kaushal et al [10], tried to create features related to videos and use them to build machine learning classifiers. The features they used were based on video keywords, comments keywords and a large list of other similar metrics. However, they did not include the actual content of the video in their feature list. With the flexibility provided by YTK to uploaders, an adversary can easily modify or poison most of these features. For example, an adversary can disable comments on their videos. This would prevent all other users from posting comments on the video in question and render all features pertaining to comments useless. Moreover, the paper also uses user-level features, which can again be forged easily. We postulate that the most prominent features for detecting inappropriate and fake content are those embedded into the video itself. A human being can easily detect an inappropriate video by just looking at its contents. Hence in contrast to [10], our system uses features directly based on the audio, video and individual frames of any uploaded material. We combine these features and create a joint deep learning architecture. More formally, our problem statement is as follows: Given a video V, we need to detect 3 labels for the video i.e. Fake F, Inappropriate I and appropriate A based on the audio-video characteristics of the video.

Our results are promising because 1. We use a deep learning based approach where no feature engineering is required. Previous studies have extracted features from videos to detect a label for them. Our approach makes use of the latest advancements in deep learning and different models like Convolutional Neural Networks [11] and Long Short Term Memory Networks [12] to build a powerful classifier that is able to detect inappropriate and fake videos with a high accuracy. All our classifier needs is a set of labeled videos. Feature engineering part is done automatically by the neural network architecture that we are using 2. Our architecture is robust against adversaries. We provide a separate section to discuss how an adversary can attack our classifier and how we have added robustness into our classifier by using different set of features instead of just using the video.

A. *Key Contributions*

We summarize our key contributions and ndings below.

- We collect a first of its kind data set that contains multiple fake and inappropriate versions of various cartoons. Multiple videos have been posted on various websites but no one has collected a single data set containing a good number of inappropriate videos before.
- Developing a deep learning based system that actually uses the content features to detect unsafe and fake videos. Our system can be extended to any type of content and can be used by YouTube itself to ag content that is unhealthy for children. We believe that this work has not been done before.

II. MOTIVATION AND PROBLEM STATEMENT

YouTube Kids Policies: YouTube Kids uses filters powered by algorithms to select videos from YouTube. However, some inappropriate content manages to seep through these filters leaving the purpose of having a separate YouTube Kids extinguished. Following excerpt from YouTube Kids sums up the current situation. [13]

We continually work hard to make our algorithms as accurate as possible in order to provide you with a safer version of YouTube. However, no algorithm is perfect ... Search and recommended videos are selected by our algorithm without human review.

Childhood exposure to inappropriate material poses serious effect on young impressionable minds. YouTubes search algorithm makes it easy for children to fall into gruesome playlist traps full of inappropriate content. Since users name their videos and use thumbnails that can get around YouTubes algorithm, a mechanism for examining the actual content of the video is crucial [14] [15]. Given this stringent issue, in this paper we present ways to tag questionable and fake videos on YouTube kids and other kids centric platforms.

Our study of the problem space gave us a notable observation: Unlike copyrighted content, only fake, copied, or derived work may contain inappropriate content. This means that if a content is not copyright of the original creators then it should be flagged. For instance, any Mickey Mouse cartoon under the copyrighted name of Walt Disney is considered appropriate, whereas cartoons containing Mickey Mouse created by someone else may or may not be safe for underage viewing. Using characters that are well sought after like Frozens Elsa and Spider-Man, YouTubers are luring children into watching offensive videos featuring their favorite characters. While at first these videos seem normal, they soon lead to those same Disney princesses and superheroes participating in lewd or violent acts.

Given these observations we divide the content into the following categories:

A. *Fake Videos*

Videos that are fake are often categorized as those which contain content not related to the entity presented in the video. Fake videos are also sometimes made from real content with slight modifications -such as facial expressions, voice and motion etc., in such a way that it closely resembles the actual content but enough to escape the copyright auto detection. In this paper we are focused on fake videos which contain such modified content.

B. *Inappropriate Videos*

There are several Video Rating Categories provided by Motion Pictures.

- **G—General Audiences:** All ages admitted. Nothing that would offend parents for viewing by children.
- **PG—Parental Guidance Suggested:** Some material may not be suitable for children. Parents urged to give parental guidance. May contain some material parents might not like for their young children.
- **PG13—Parents Strongly Cautioned:** Some material maybe inappropriate for children under 13. Parents are urged to be cautious. Some material may be inappropriate for pre-teenagers.
- **R—Restricted:** Under 17 requires accompanying parent or adult guardian. Contains some adult material. Parents are urged to learn more about the film before taking their young children with them.
- **NC 17—Adults Only:** No One 17 and Under Admitted. Clearly adult. Children are not admitted.

For our analysis, videos that fall under the G General Audiences Category are considered appropriate while videos in rest of the categories are considered inappropriate. However, this content may be real or fake. The inappropriate videos either contain a) Explicit Videos or b) Violent Videos. We label this inappropriate content **I** and all other content is labelled **A**.

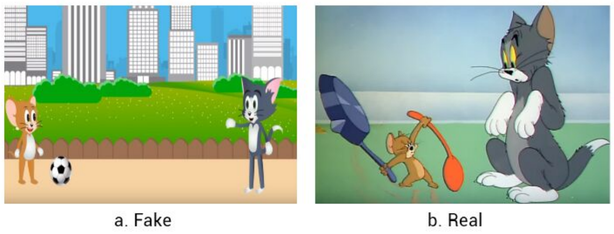
Problem Statement

More formally, the categories can be labelled as follows:

- V = all content i.e. the entire sample space.
- R = real content or content made by the actual creators.
- F = fake content.

$$F = V - R$$

- A = appropriate or child safe content.



a. Fake

b. Real

Fig. 1. Contrast of Fake and Real Tom and Jerry cartoons.

- I = inappropriate content (both Explicit Videos as well as Violent Videos).

$$I = V - A$$

Real videos are mutually exclusive to inappropriate videos:

$$R \cap I = \otimes$$

III. DATA

A. Data Collection

Our data is composed of four sets of videos, the first contains original cartoon videos collected from their official channels on YouTube, the second set contains fan made or fake videos also obtained from YouTube, and the third set contains explicit videos obtained from various sources.

*1) **Original Videos**:* We obtained this set from official cartoon channels from YouTube. This set contains videos of Peppa Pig, Mickey Mouse, and Tom and Jerry etc. uploaded under the name of their official copyright owners and is deemed real. These videos were downloaded and watched to ensure that the content is appropriate.

*2) **Fake Videos**:* For the same cartoons (i.e. Peppa Pig, Mickey Mouse and Tom and Jerry) we collected fan-made and fake cartoons on YouTube. This difference can be observed in Figure 1. The characters in the fake videos are rendered very differently, their movement is jerky and lacks originality. Most of the fan-made and fake cartoons are usually uploaded by a single channel, hence reducing our search space. We proceeded further by downloading all the videos of such channels.

*3) **Explicit Videos**:* We first obtained an initial pool of explicit videos by combining the names of famous cartoons with an explicit keyword such as kiss, pregnant, bikini, love story, etc. A few months ago, a post was submitted on Reddit with the title *What is up with the weird kids videos that have been trending on YouTube lately?*[16] .The article has received 4k up votes till now. Here we found a huge set of actual videos on YouTube that are explicit and parents have deemed unsuitable. Further rigorous digging into this subject directed us towards articles and blogs that have voiced this issue and also mentioned various videos that were inappropriate for children.

Secondly, while the aforementioned videos were being analyzed, a second pool of such videos were collected that were typically suggested and recommended by YouTube as a result of watching these videos. Usually the same channel uploaded more explicit videos.

Data Set	Videos Count
Original Cartoons	487
Fake Cartoons	546
Explicit Videos	213
Violence Videos	150
Total Videos	1396

TABLE I
VIDEOS PER CATEGORY

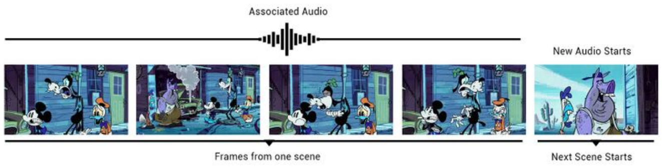


Fig. 2. Segmented Scene

*4) **Violent Videos**:* Similarly, to obtain the set of violent videos, we queried YouTube with phrases containing combinations of cartoon character names and keywords that specifically show violence: for example knives, blood, and violence etc. Two such cartoons that were found to be violent are Scottish Ninjas and Happy Tree Friends. Furthermore, we found a particular YouTube channel Mondo Media that publishes violent cartoons like Rig Burn and Zombie College. Each of these videos were manually tested and analyzed with the time frames containing violence being marked.

Table 1 lists the exact number of videos we collected in each category.

IV. DATA PROCESSING

Video Classification using machine learning is a time intensive task, since a single video can contain many frames. Therefore, before we jump into applying deep learning to classify data, we process the collected data:

1) Scene Segmentation: Each video is converted into various sets of clips. Figure 2 shows an example of a segmented scene. Each clip constitutes a complete scene. To do so we use image histograms of video frames. This gives us an initial array for red, green and blue colors which are later appended to form a single array for each frame. Each histogram is compared with the mean of the previous three histograms using cosine distance. Upon observation of the cosine distances with threshold 0.4, we concluded that every scene change results in spikes. Thus we define a threshold T for this distance; whenever the distance becomes greater than T, we begin a new clip and save the previous. For our system we have set T = 0.15, however, T can be considered as an input parameter, and must be adjusted accordingly.

Furthermore, many of the scenes are very short which results in multiple spikes within a few frames.

To incorporate this we only keep scenes that contain more than a certain number of frames.

Choice of Threshold: The value of threshold T, depends on the contrast, saturation and brightness of the video. We have adjusted the value of T using trial and error based on two major cartoon settings. The first setting includes cartoons with similar colors throughout their videos such as Smurfs

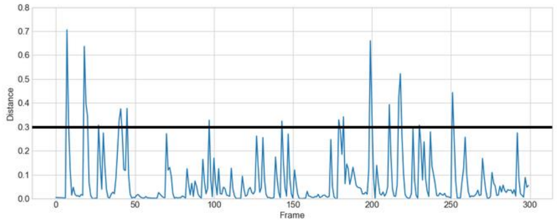


Fig. 3. Distance Plot for a Smurf episode. The spikes are close to 0.3 which is lower than Tom and Jerry graph.

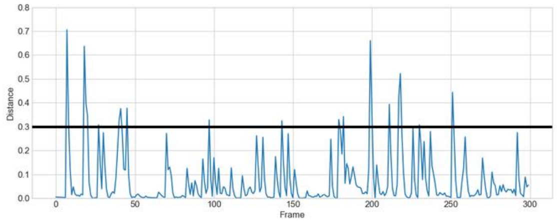


Fig. 4. Distance Plot for a Tom and Jerry Episode. The spikes on scene change are always close to 0.4

and Octonauts episodes. The second setting includes cartoons with contrasting colors. For this, we choose Mickey Mouse and Tom and Jerry episodes. Analysis of plots of frame distances showed that in the first setting the spikes (distance from previous frames) is relatively low, whereas in the second setting the spikes are usually high when the scene changes. Figure 3 and 4 show plots of frame distances of the two settings.

2) Audio Segmentation: For each video clip segmented in the previous step we extract its audio and convert it into a spectrogram to be used by our deep learning module. Figure 5 shows an example of the produced spectrogram.

A. Feature Extraction

To extract features from our frames and spectrograms we use a pre-trained VGG-19 [17] convolutional neural network. Using the output of the third last layer of the CNN as our feature vector we get a 4096 dimensional array as output. For each scene, there are three sets of features that we extract from the CNN.

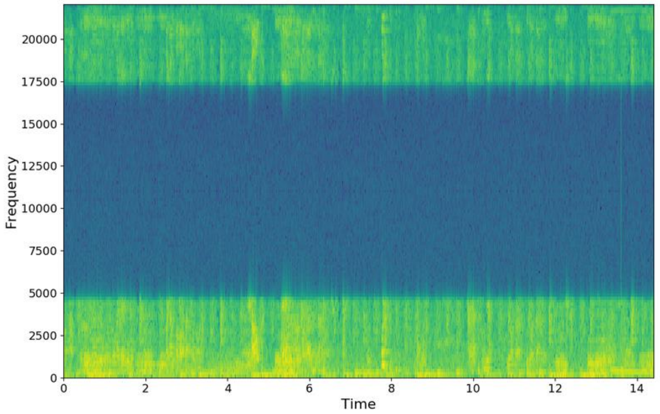


Fig. 5. Audio Spectrogram

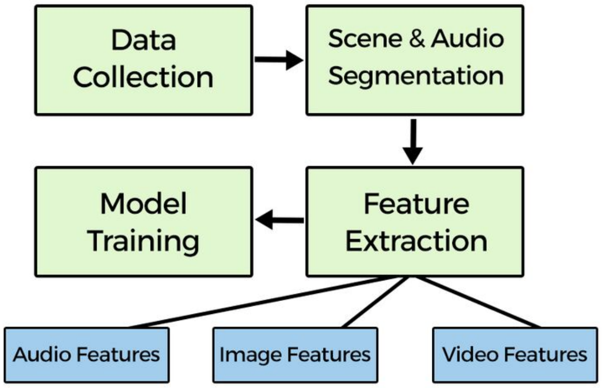


Fig. 6. System Workflow

1) Frame Features: To extract features of a frame, we pass it through a pre-trained CNN. Instead of using the entire cartoon we segment it into frames to easily recognize any fake frames if amalgamated in the cartoons. Moreover, frames of fake videos are rendered differently, and usually, if a single frame in a cartoon contains explicit content, we can mark the entire cartoon as inappropriate.

2) Movement Features: The difference between the movement of fake and real cartoon characters is quite significant. In a fake cartoon the characters hardly move their limbs. Their movement is jerky and rigid; therefore, features related to character movement are pivotal for our model. Movement features are extracted after each frame is passed through the CNN and saved in a 2-Dimensional array of size n x 4096 where n is the number of frames in a scene.

3) Audio Features: To extract audio features, each frames audio spectrogram is passed through the CNN and its output recorded. Almost all fake and explicit videos have little or no audio. Usually, music plays in the background, but the characters themselves do not speak anything. Conclusively, audio features are an important factor in distinguishing inappropriate videos.

V. SYSTEM WORKFLOW

After we segment each video into different segments, each segment (constituting of multiple frames and audio spectrograms), is passed through a Convolutional Neural Network (CNN), to extract its features. All the features are later fed into a deep learning system, where we perform training to predict the label. Figure 6 shows the workflow of our system.

1) Feed Forward Neural Networks: Figure 7 shows a basic 3 layered neural network [18]. The first layer in the figure represents the input layer. The input features are projected onto this layer. The middle layer of nodes is called the hidden layer, because its values are not observed in the training set. Each

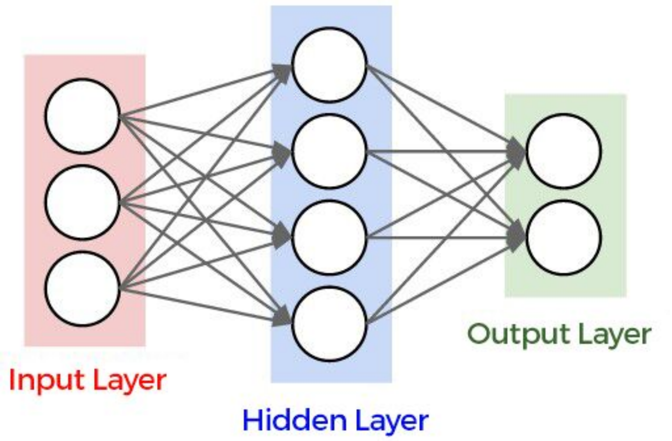


Fig. 7. Feed Forward Neural Network

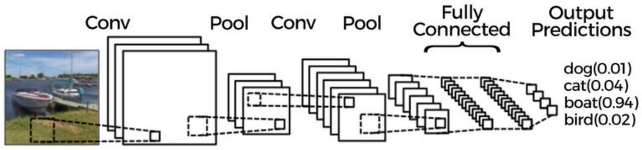


Fig. 8. Convolutional Neural Network

node in this layer is a weighted sum of input layer features. The final layer gives out the output of the network. Each node in the output layer presents the probability of a certain class of output. Information in the form of features is propagated forward in the network during forward feed and the error in the final layer output is then used to adjust weights in the network during back propagation phase. A neural network performs many iterations of forward feed and back propagation during execution.

2) Convolutional Neural Networks: Convolutional Neural Networks [11] are mostly used for analyzing and performing tasks related to visual imagery. The primary purpose of Convolution in our project is to extract features from the input frames. Spatial relationship between pixels is preserved by learning image features using small squares of input data during Convolution [19]. As shown in the Figure 8, in a Convolutional Neural Network a filter is convolved over the entire image. Images are usually represented as an array of shape (height, width, channels), where there are 3 channels for colored images and 1 for grayscale images. For a filter to be able to convolve over the image, it should have the same number of channels as the input image. The output is the dot product of the filter elements and the image. Every dot product yields a scalar output, therefore after every convolution the original size of the image decreases. Conclusively, a neural network learns weights and a CNN learns these filters.

3) Recurrent Neural Networks: Recurrent nets are a powerful set of articial neural network algorithms especially useful for processing sequential data such as sound, time series (sensor) data or written natural language. Recurrent neural nets differ from feedforward nets because they include a feedback loop where the output of the input depends on the current time step as well as the previous time steps. Figure 9 represents a basic RNN. Each cell computes the output as well as transfers it to the next stage. In our model, we a use a special type of RNN called Long Short Term Memory Network (LSTM) capable of learning long-term dependencies [12]. It has the ability to remove or add information to the cell state, carefully regulated by structures called gates.

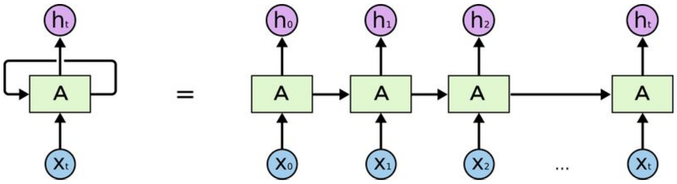


Fig. 9. Recurrent Neural Networks

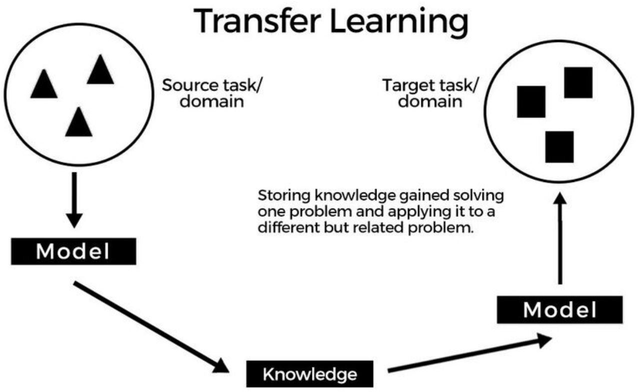

Transfer Learning
Source task/ domain
Target task/ domain
Storing knowledge gained solving one problem and applying it to a different but related problem.
Model
Knowledge
Model

Fig. 10. Transfer Learning

4) Transfer Learning: Transfer Learning [20] is a Machine Learning Problem that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. In the past decade, transfer learning has been applied effectively when transferring knowledge across domains [12]. Figure 10 shows two different tasks with common domain knowledge. A model is trained on the source task and since the domain of both the tasks is same, the knowledge obtained from the trained model can be used for training the model for Target Task.

5) Model Overview: According to Figure 11, our model takes segmented frames and audio spectrograms of a single scene as input. These inputs are passed through a pre trained CNN to extract features. Extracted frame and audio spectrogram features are kept the same, however, features related to character movement are further preprocessed using Recurrent Neural Networks (RNN) to extract the precise movement level features of the specific scene. The output of RNN is a 1-Dimensional array which is concatenated with the audio and frame level features. The neural network then outputs the label of the scene after passing it through the dense layer.

VI. EXPERIMENTS

We begin by using the three features separately for classification, and later proceed to see whether the combination of

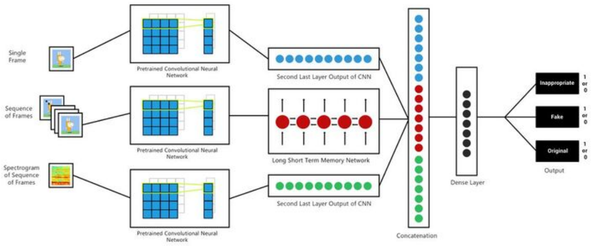

Single Frame
Sequence of Frames
Spectrogram of Sequence of Frames
Pretrained Convolutional Neural Network
Pretrained Convolutional Neural Network
Pretrained Convolutional Neural Network
Second Last Layer Output of CNN
Long Short Term Memory Network
Second Last Layer Output of CNN
Concatenation
Dense Layer
Inappropriate
1 or 0
Fake
1 or 0
Original
1 or 0
Output

Fig. 11. Model Overview

Method	Accuracy	Precision	Recall
Audio	66.6	66	67
Frame	67.7	69	62
Movement	83.6	83	84
Architecture	88.24	88	88

TABLE II
RESULTS FOR NON-EXPLICIT VS EXPLICIT VIDEOS. NUMBERS ARE ROUNDED UP FOR PRECISION AND RECALL.

the three will improve results. The data-set is split into a ratio of 80:20, where 80% of the data is used as training set and the rest 20% is used for testing. To obtain consistent results, we also apply 5 fold cross validation for each experiment. The scores reported in our results are average across all 5 folds.

A. Analysis Using Individual Features

1) Audio Features: The output from the CNN for audio spectrogram is passed to the feed forward neural network and its accuracy measured. Since there are clear differences between the audios of original and fake/explicit videos we expect high accuracy using only audio features.

2) Individual Frame Features: Similarly, we pass the features for individual frames picked randomly from the scenes to a feed forward neural network and measure its accuracy.

3) Movement Features: We pass our movement features to RNN followed by a feed forward neural network and then measure the scores.

B. Analysis using Combination of the Features

We experiment by combining all the features and passing them to our joint deep learning architecture. We expect that the combination of the three features will increase the predictive power for the neural network, because for each scene there will be multiple factors contributing towards the prediction of the label. The joint architecture is created in Keras [21]. FC1 layer of VGG19 has been used to extract image features. The model is run for 25 epochs with a learning rate of 0.00001.

VII. RESULTS

Table 2, 3 and 4 show the results obtained after performing training using individual features as well as after using a combination of the three features. Even though individually each features performs good, however, the best results are obtained after the three features are used in combination.

Method	Accuracy	Precision	Recall
Audio	89.3	89	90
Frame	65.9	75	64
Movement	88.4	89	89
Architecture	96.01	96	96

TABLE III
RESULTS FOR REAL VS FAKE. NUMBERS ARE ROUNDED UP FOR PRECISION AND RECALL.

Method	Accuracy	Precision	Recall
Audio	95.05	95	95
Frame	62.31	69	61
Movement	81.54	82	82
Architecture	95.9	95	95

TABLE IV
RESULTS FOR NON-VIOLENT VS VIOLENT. NUMBERS ARE ROUNDED UP FOR PRECISION AND RECALL

As we can see from table 2, audio analysis for non-explicit vs explicit videos does not give higher accuracy since most of the explicit videos we analyzed did not have distinct audio. Although some videos had kissing noise, not all of them had enough audio to help our classifier classify such videos. Furthermore, contrary to our expectation, using frames individually to classify explicit videos is not enough, and the accuracy remains low. This is because we randomly picked frames from a video and might have missed an explicit scene from a video that overall looks appropriate. However using the entire video and using content movement and sequence as feature yielded the highest results. Using sequence of scenes to detect an explicit act is intuitive, since our model can learn when characters are moving in to kiss etc. Conclusively the joint architecture outperforms all the others.

Most of the fan made and fake videos use very simple audio. They use background music where the characters hardly talk. Therefore, using audio as a feature to differentiate between real and fake videos gives a high accuracy. However, similar to the explicit vs non-explicit results, our image analysis does not give us a considerably high accuracy due to random picking of scenes. Not all the frames inside a scene are able to distinguish a real video from a fake one. Using sequence of frames and content movement worked better, however, joint architecture gave us the maximum accuracy. This ablation analysis gave us an interesting insight that audio from video is sufficient to differentiate between fake and real videos.

Figure 12 shows the change in accuracy due to increasing the number of iterations. Test accuracy becomes constant as the number of iterations exceed 10. We are avoiding overfitting by using high dropout value (0.5).

Throughout the paper we mentioned classification of individual scenes rather than the complete video. We do this to maintain robustness of the classifier. Since we are performing scene-wise classification, an adversary cannot perturb the video to misclassify it. We discuss adversaries further in the paper. However, we propose a simple algorithm to classify a complete video. Algorithm 1 describes method to classify a video.

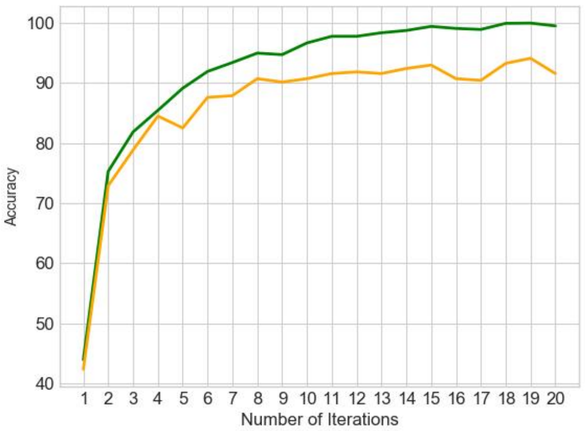


Fig. 12. Accuracy increase with the increase in number of iterations

Algorithm 1 Algorithm to classify complete video

```
**Data:** Video V
**Result:** Label L for V
Output is 1 if the video belongs to L and 0 otherwise
Set a threshold X
Segment V into scenes
$TotalScenes\ = 0$
$ClassiedAsL\ = 0$
**for** each scene S **do** $ClassifedLabel\ = classify(S, L)$
    **if** $ClassifedLabel\ == 1$ **then**
        $ClassiedAsL\ = ClassiedAsL + 1$
    **else**
        $TotalScenes\ = TotalScenes + 1$
    **end if**
**end for**
$ClassiedThreshold\ = ClassiedAsL/TotalScenes$
**if** $ClassiedThreshold\ ! = X$ **then**
    Video is labeled as L
**else**
    Video is not labeled as L
**end if**
```

VIII. Case Studies

In this section, we present a discussion on the performance and limitation of our model. For this purpose, we will classify videos belonging to different categories from the test set.

A. Appropriate and Normal Video

We analyzed three different cartoons to demonstrate the performance of our classifier on a normal video. The first video was from the cartoon Tom and Jerry, the second one from Peppa Pig and last one from Popeye. In order to classify the videos, we randomly extracted scenes from them. For each scene in the videos, the number of frames was fixed to 15. The table 5 illustrates the results that we achieved.

An evidence of our classifiers ability to correctly label unseen data is that there were no videos from Popeye in our training set. Yet, the classifier achieves 100% accuracy in labeling its scenes correctly.

Cartoon	Scenes	Classification Accuracy
Tom and Jerry	9	100%
Peppa Pig	15	100%
Popeye	15	100%

TABLE V

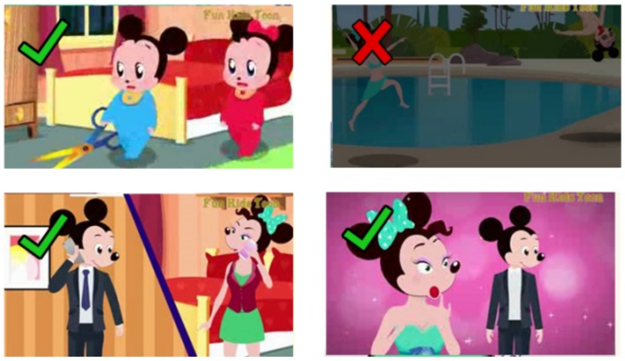


Fig. 13. Frames from a Fake Mickey Mouse Video. 87% of the frames were labeled correctly by our classifier. Only 2 scenes are labeled incorrectly with high probability (>70%)

B. Appropriate and Fake Video

We also examined videos which were appropriate albeit fake. For this purpose, we ran our classifier on 15 scenes from a fake Tom and Jerry video. Our model labeled 10 scenes correctly with high probability (>90%). From the misclassified scenes, 1 was labeled with high probability whereas the remaining 4 were assigned almost equal probabilities to the likelihood of the scenes being fake or real.

To ensure consistency in our results, the model was run on a fake Mickey Mouse video as well. Out of the 15 scenes, 13 were labeled correctly and 2 were labeled incorrectly. As evident in Figure 13, one of the incorrectly labeled scenes was dark, which might have led to its erroneous classification.

C. Explicit Video

We examine three different videos with distinct explicit scenes to provide a comprehensive analysis of suggestive videos. The videos used for testing contained intimate scenes and inappropriate dressing.

Our first test video contains women wearing revealing clothes and engaging in salacious activity. The entire content of the video is explicit and our classifier is able to detect all 11 scenes correctly with a high probability (>99%). There is one scene which is incorrectly labeled as explicit (evident in Figure 14). However, the probability of it being explicit according to our model is only 50.5%.

As our next experiment, we ran the classifier on an explicit video of Mickey Mouse which only had one inexplicit scene. All of the scenes were correctly labeled by the classifier.

In our last test video, the classifier misidentified a scene as non-explicit (as shown in Figure 15). The remaining 14 were

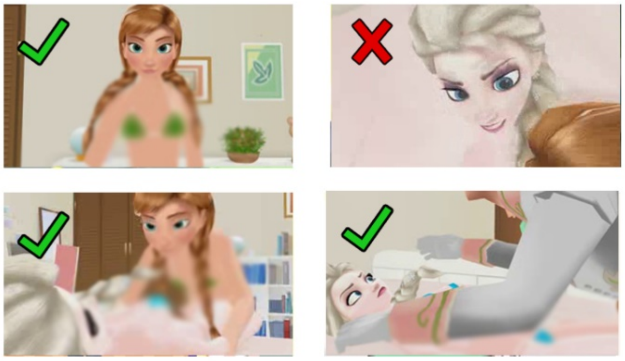


Fig. 14. Frames from an Explicit Video. 100% are labeled correctly. The scene with red-cross was not explicit based on its frames but the model labeled it as explicit because of its audio.

Fig. 15. Frames from an Explicit Video. 93% of the frames were labeled correctly by our classifier. The red-cross denotes that the scene is non-explicit which is incorrect.

labeled correctly.

For complete experimentation, we also tested our model on a video containing violent scenes. The results obtained are represented in Figure 16.

A hug is a great gift - one size fits all and it's easy to exchange!

Fig. 16. Frames from a Violent Video. 83% of the frames were labeled correctly by our classifier. The red-cross denotes that the scene is not violent. It was also detected accurately by our classifier.

IX. Adversarial Analysis

In the above case study we took videos in their original form and analyzed them. However, in this section we will be discussing adversarial analysis; what happens if an adversary perturbs a video or some of its features. More precisely, we will remove some of the features of the video; a type of ablation analysis and test it using our joint architecture.

1) Frame Replacement: An adversary can try replacing random frames from an explicit or violent video, with appropriate and non-explicit ones. However, such an attack will fail in our system, because during feature extraction we pick frames randomly out of a scene. The adversary can never know the exact frame that our model picks from a specific scene. To quantify our claim, we replaced a single frame in all scenes of an explicit videos with a frame from an appropriate video. To our expectations, the accuracy remained the same and the model was able to detect all inappropriate scenes of the explicit video.

2) Audio Replacement: Our joint architecture does not rely on audio alone, hence, such an adversarial attack will not pass through our system, unnoticed. To test our claim, we replaced the audio of each scene from an explicit video with audio from original Tom and Jerry. The accuracy remained the same and the model was able to detect all explicit scenes correctly.

3) Scene Replacement: The adversary can try replacing random scenes of a video. Here the adversary can add appropriate and non-explicit video clips at random locations of a video. However, this too wont work on our model, because instead of labeling the entire video, we label scenes as explicit/non-explicit, violent/nonviolent or fake/real. To test this, we randomly added clips from an original and appropriate cartoon in an explicit video and tested our model. The model accurately detected all the explicit scenes and the appropriate ones that were added were not labeled by our system.

X. Discussion and Conclusion

Since, YouTube became prevalent among kids, a separate mobile application YouTube Kids in response to anxiety from parents, was launched. However, this resulted in children being advertised to fake, inappropriate content directly [22].The YouTube Kids app extends the reach of YouTube into the lives of kids. This paper presents a deep learning architecture to detect fake and inappropriate videos that are targeted for kids. We collect a first of its kind data set for inappropriate videos and develop a system to flag such videos. Our results show that this approach can be successfully applied on several cartoons to detect inappropriate content in them.

XI. Related Work

Video forgery Detection is an active area of research with people using different techniques for video retrieval and fingerprinting. C. Wu et al. [11] performed content based video classification by representing each frame as randomly projected binary features and using them to measure similarity between videos. M. Esmaeili et al. [12] used four types of video ngerprints (color-space, temporal, spatial and spatiotemporal) to find closest match for each video in the database. These features were extracted from temporally informative images from video sequences and then a 3D-DCT algorithm was used to perform spatial-temporal fingerprinting for video matching on them.

Yuan et al. [20] developed a robust transformation-invariant video fingerprint using natural parts (coarse scales) of the Shearlet coefficients for revealing spatial and directional features of a video and used TRECVID dataset to exhibit its strong ability of discrimination and robustness for variety of video copy attack. Wahab et al. [23] surveyed three types of video forensic techniques including statistical correlation of video feature, frame-based techniques for detecting statistical anomalies and inconsistency features of different digital equipment.

Different match learning techniques are also used for capturing subtle differences in video frame. Avino et al. [17] proposed a method for video splicing for detection based on auto-encoder and recurrent neural networks. They used image patches and extracted handcrafted features from them. This captures subtle statistical differences between spliced materials with respect to original video. The feature vector was then used in auto encoder for anomaly detection.

Object-based forgery detection has also been applied to detect fake videos. It involves by first detecting forged objects in a video and comparing them to the actual video objects. Chen et al. [17] proposed an algorithm to perform object-based video forgery detection. The algorithm first performs frame manipulation detection using image forensic methods to find tampering in motion residuals and then uses a two stage algorithm to locate the coarse boundaries of forged segment in the frame and fine tune them. Sunil Lee et al [4] proposed binary fingerprinting using a feature selection algorithm called the symmetric pairwise boosting (SPB). The algorithm extracts features from a preprocessed video, and then chooses appropriate filters and quantizers, for boosting the performance of binary features. These features representing binary fingerprints of videos are used for content matching. The perceptually similar and dissimilar pairs of video clips are correctly classified as matching or non-matching pairs using these fingerprints.

Bekhet and Ahmed [6] used Dominant Color Profiles to produce a set of DCPs across video acting as a compact signature for the entire video. Every block of a DCP image is represented by its dominant color (spike), where the sequence of dominant colors for each block is kept as descriptive color profile. Both spatial and temporal information are captured inside DCP which allows for real time content matching of videos. Results showed DCPs robustness in real time video matching and efficient computation.

XII. Future Work

Future applications of this project can involve analyzing what the scene is about. We aim to expand our project further by introducing methods to extract what is actually happening in the scene. Up till now we label the scene as explicit, violent, or fake. However, further work can be done, to let parents know what is exactly wrong with a video. For example, if a cartoon video contains kissing, we can not only label the video as explicit, but also tag that kissing is involved in the video.

References

[1] "We spend insane amounts of time watching youtube on phone." https://www.wired.com/2015/07/spend-insane-amounts-time-watchingyoutube-phones, July 2015.

[2] "How much time will the average person spend on social media during their life? (infographic)." http://www.adweek.com/digital/mediakixtime-spent-social-media-infographic, March 2017.

[3] "Kids under 8yo spend 65 percent of their online time on youtube." https://www.familyzone.com/blog/what-kids-did-online-2016, January 2017.

[4] "The disturbing youtube videos that are tricking children." http://www.bbc.com/news/blogs-trending-39381889, March 2017.

[5] "Youtube kids features some pretty disturbing content." https://thenextweb.com/insider/2017/03/27/youtube-kids-features-prettydisturbing-content/, March 2017.

[6] "How elsa, spider-man trick kids into watching violent youtube videos." http://mashable.com/2017/10/22/youtube-kids-app-violent-videos-seokeywords, October 2017.

[7] "On youtube kids, startling videos slip past filters." https://www.nytimes.com/2017/11/04/business/media/youtube-kidspaw-patrol.html, November 2017.

[8] "Advocates charge google with deceiving parents about content on youtube kids, request ftc action." http://commercialfreechildhood.org/advocates-charge-google-deceivingparents-about-content-youtube-kids-request-ftc-acti May 2015.

[9] "Youtube kids app reported to ftc for featuring videos with adult content." https://techcrunch.com/2015/05/19/youtube-kidsapp-reported-to-ftc-for-featuring-videos-with-adult-content, May 2015.

[10] Kaushal and R. et al, "Kidstube: Detection, characterization and analysis of child unsafe content and promoters on youtube," *Privacy, Security and Trust (PST), 14th Annual Conference on IEEE*, 2016.

[11] Krizhevsky, Alex, I. Sutskever, and G. E. Hinton, "Imagenet classication with deep convolutional neural networks," *Advances in neural information processing systems*, 2012.

[12] Hochreiter, Sepp, and J. Schmidhuber, "Long short-term memory," *Neural computation 8.9 1735-1780*, 1997.

[13] "Youtube kids parental guide." https://support.google.com/youtubekids/answer/6130561?hl=en.

[14] S. Maheshwari, "On youtube kids, startling videos slip past filters." https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html, November 2017.

[15] J. Alexander, "Youtube kids has been a problem since 2015 — why did it take this long to address?." https://www.polygon.com/2017/12/8/16737556/youtube-kids-video-inappropriate-superhero-disney, December 2017.

[16] "Loop." https://www.reddit.com/r/OutOfTheLoop/comments/65rkxv/what_is_up_with_the_weird_kids_videos_that_have/.

[17] "Moms warn of disturbing video found on youtube kids: Please be careful." https://www.today.com/parents/moms-warn-disturbing-video-foundyoutube-kids-please-be-careful-t101552, November 2017.

[18] Hornik, Kurt, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks 5.2, 359-366*, 1989.

[19] "An intuitive explanation of convolutional neural networks." https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/, August 2016.

[20] Oquab and M. et al, "Learning and transferring mid-level image representations using convolutional neural networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[21] "Keras, the python deep learning library." https://keras.io/.

[22] B. Burroughs, "Youtube kids: The app economy," *Sage*, 2017.

[23] Karpathy and A. et al, "Large-scale video classification with convolutional neural networks," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014.