

Bringing the Kid back into YouTube Kids: Detecting Inappropriate Content on Video Streaming Platforms

Rashid Tahir*, Faizan Ahmed†, Hammas Saeed‡, Shiza Ali‡, Fareed Zaffar‡ Christo Wilson§

*University of Prince Mughrin, KSA, r.tahir@upm.edu.sa

† University of Virginia, USA, fa7pdn@virginia.edu

‡ Lahore University of Management Sciences, PK, ([hammas.saeed](mailto:hammas.saeed@lums.edu.pk), [shiza.ali](mailto:shiza.ali@lums.edu.pk), [fareed.zaffar](mailto:fareed.zaffar@lums.edu.pk))@lums.edu.pk

§ Northeastern University, USA, c.wilson@northeastern.edu

Abstract—With the advent of child-centric content-sharing platforms, such as YouTube Kids, thousands of children, from all age groups are consuming gigabytes of content 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 of inappropriate videos hosted on such children-focused apps and platforms. Alarming, our study finds that there is a noticeable percentage of such videos currently being watched by kids with some inappropriate videos having millions of views already. To address this problem, we develop a deep learning architecture that can flag such videos and report them. Our results show that the proposed system can be successfully applied to various types of animations, cartoons and CGI videos to detect any inappropriate content within them.

I. INTRODUCTION

YouTube has become so ubiquitous among kids that according to a recent study [1], 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, 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 [2], inappropriate content often lands on YTK. From videos containing references to sex and alcohol to drug use and pedophilia, a wide range of unsuitable content has been reported on YTK [3]. The prevalence of these videos

has recently gained more traction with parents and many news outlets have raised concerns over the widespread problem [3]. 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 falsely marketing YTK as being safe for children [4].

Of course, the intent behind the YTK app is noble and Google has deployed numerous automatic checks, such as filtering algorithms, and manual review mechanisms to filter out unwanted content. However, due to the sheer volume of the content, manual review is not always possible. Furthermore, given past trends of the rate at which videos are being uploaded, manual review will become impractical soon (if it hasn't already). Hence, we assert that improvements need to be made to the auto-filtering mechanisms by leveraging deep learning methods, which will allow flagging of inappropriate content that can then be reviewed manually or passed through additional checks.

To this end, researchers have tried to propose various solutions. For instance, authors in [5] 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 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 corresponding to comments useless. 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 [5], we propose a system that uses features directly based on the audio and video properties of any uploaded material. We combine these features and create a joint deep learning architecture.

Our main contributions are the following: First, we manually watched around 5k videos (in entirety) to collect a large-scale dataset containing unsuitable content on YouTube Kids, spanning fake cartoons and animations to inappropriate rip-offs and advertisements. Small samples have previously been

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ASONAM '19, August 27-30, 2019, Vancouver, Canada

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6868-1/19/08/\$15.00

<https://doi.org/10.1145/3341161.3342913>

posted on various websites but no one has made a concentrated effort to create a single dataset containing a decent number of inappropriate videos before (to the best of our knowledge). Our data set includes more than a 1000 videos covering fake, explicit or violent content. Second, we add-on to existing techniques by developing a deep learning-based system that can be used to flag the aforementioned unsafe content. Using our classifier on a hold-out sample during cross validation, we manage to achieve an accuracy of more than 90%.

II. PROBLEM OVERVIEW

Online video streaming has become commonplace in our times. However, for platforms dedicated to children major concerns revolve around the presence of sexually explicit language, pornographic material, profanity, pedophile jokes and child-unsafe behavior. Disturbing videos are being uploaded to these platforms (YouTube Kids and others), leveraging popular cartoon characters, such as Elsa from Frozen, Spiderman, Snow White, etc. , without being flagged [6]. These favorite characters are first used to lure children in, using catchy thumbnails and misleading titles and then misused to expose the young audience to sexually explicit and violent content [7]. Similarly, pranksters from the *4chan* group were able to manipulate YouTube’s algorithms and place pornographic content into children’s shows [8], where the targeted videos started off innocently but eventually led to explicit scenes. The problem is exacerbated by the sheer amount of content, which makes it impossible to ensure manual human review for *all* the content (around 400 hours of video data is uploaded every minute on YouTube [9]) and we feel that the algorithms that are currently in place can be improved.

Harmful Effects of Unsafe Content Consumption: Child psychology experts assert that exposure to such content can possibly desensitize children to violence and they might reenact what they see at homes or schools [7]. It has also been argued that exposure to inappropriate videos can cause serious behavioral modifications among young impressionable children [10]. Another alarming notion is that children grow to “know” and “trust” some of these characters and develop an affinity towards them. When these characters are involved in violent and improper acts, this is greatly upsetting for the children and creates frustration and anxiety in them [2].

Exploiting YouTube’s Algorithms and Features: Content producers on YouTube have devised various methods to draw attention to their videos. To this end, they sometimes use a two-phased approach with various tricks and tactics leveraged in each phase. First, they try to trick the search algorithms and “squeeze” inappropriate content into the result of a legitimate query. Second, they lure the child into opening one of the target videos using various baiting strategies. Below we explain some of these tricks that uploaders use during each phase.

To begin, uploaders typically use “word salads”; a big hotchpotch of words including names of trending characters, toys, nursery rhymes etc., to maximize chances of a hit [11]. A simple search for “Peppa Pig” for instance, contains results

from channels other than the verified channel, even on the first page of search results. This is typically achieved by using the word “Peppa Pig” in the title and the associated metadata. Another common strategy involves compiling cartoon videos, rhymes and other content together into a long video [11]. These videos typically contain scenes not suitable for children e.g., involving the main character killing people, bullying or showing explicit behavior etc. Additionally, YouTube’s search algorithm also makes it easy for children to fall into playlist traps full of inappropriate content (streaming one video after another via the auto-play feature). In such a case, the “Recommendations” feature by YouTube becomes a tool for abuse as a single inappropriate video can lead to similar suggestions, often from the same channel. Since children are inherently curious, it is quite likely for them to get deceived and open an incorrectly recommended video [7]. Other popular forms of baiting include content similar to popular songs/rhymes from children movies or catchy thumbnails depicting popular characters doing something the children have not seen them do before (like Spiderman in a different costume). We have found numerous videos in our datasets with catchy thumbnails that have thousands of views despite having no meaningful content in them. More advanced users have also been found relying on bots to push their videos up in the search results by artificially inflating the popularity of their content [12]. For instance, newly created videos are often given an initial “push” by having a large number of bots view the video repeatedly.

In light of the aforementioned challenges, we postulate that a scalable detection mechanism that “dives” into the actual contents of the videos rather than the titles or keywords is needed to perform any meaningful filtering. Hence, in this paper, we present ways to tag questionable and fake videos on YouTube Kids and other child-centric platforms. We make the following observation: Unlike original copyrighted content, fake, copied, or derived work typically contains inappropriate content (whether sexually explicit or content containing violence etc.). This is primarily because copyright material creators can be held accountable (like Disney) and hence, act in a responsible way so as to avoid backlash. Fake or copied work, on the other hand often means limited or no liability. Hence, we postulate that if content is not copyright of the original creators (it is fake) then it should be flagged and inspected further.

Given these observations we divide our target content into the following three categories of fake videos:

1. Benign Fake Videos (BFV): Videos that contain content *associated* with an existing copyright but not created by the original copyright owners are categorized as *BFV*. These videos are made from real content with slight modifications, such as facial expressions, voice and motion changes etc., in such a way that it closely resembles the actual content but different enough to bypass YouTube’s copyright detectors. Despite being fake, these videos do not include explicit or violent content (e.g., videos with subliminal advertisements).

2. Explicit Fake Videos (EFV): We rely on the *Video Rating System* designed by the Motion Picture Association

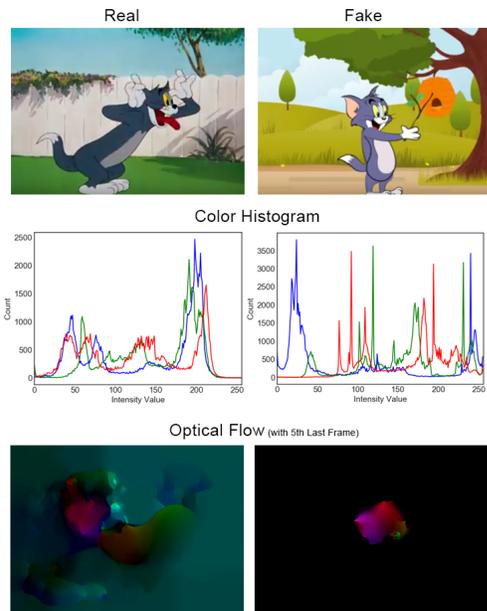


Fig. 1: Comparison of fake and real versions of the popular Tom and Jerry cartoons. We can see a difference across various parameters in the fake and real versions.

of America (MPAA) to establish benchmarks for inappropriate videos (explicit category). Content that does not contain anything offensive or explicit for any age group is rated **G – General Audiences** by the MPAA. Parental supervision is not needed as the category is deemed appropriate for viewing by children. *Toy Story* (1995) for instance, is rated *G*. For our analysis, videos that fall under the constraints of category *G* are considered appropriate while videos in all other categories (such as *PG*, *PG13*, *R*, *NC17*) are considered *explicit*. Once a fake video is flagged, an MPAA-rating is assigned to it to determine if it falls under the explicit category or the aforementioned BFV category.

3. Violent Fake Videos (VFV): The last category comprises fake videos that contain violent scenes such as those involving blood, gore etc. We specifically look for themes that revolve around killing, beating, fighting etc. Again we follow the guidelines provided by the MPAA rating system. If a video falls under any category other than *G* due to depictions of violence then we categorize it as violent (VFV).

III. DATA COLLECTION

We manually watched around 5k videos to curate the dataset, which was a painstaking process requiring a substantial investment of time. A team of 3 members ran hundreds of queries/searches to see the responses and spent around 300 hours of binge watching videos (chaining from one video to another). We believe that the end product is a dataset that is representative of the real world. However, we do concede that due to the ever-evolving nature of search results (changing popularity and ratings, viral videos, algorithmic changes from Google and whole host of unpredictable parameters) along with a staggering scale at which new videos are being uploaded, our dataset has probably become stale. Nevertheless,

in terms of providing a labeled collection of inappropriate and fake videos harvested directly from YTK, it still has value. For instance, the dataset provides insights into what tricks and tactics are used by abusers, which age-group is targeted the most, what are easier avenues of detecting said abuse etc., and hence, it serves to benefit the community. Furthermore, it is large enough to train a classifier in a meaningful way and hence, allows researchers to try different approaches.

Our dataset consists of three sets of videos. **Original videos** were obtained from official cartoon channels on 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 were manually labeled to ensure appropriate content. For the same cartoons, we collected a sample of fan-made and **fake videos**. The difference between a real and fake cartoon (tom and jerry) can be observed in Figure 1. The characters in the fake videos are rendered differently, their movement is jerky and lacks originality. To show this, the second image in Figure 1 shows the colored histograms. The histograms of real tom and jerry are flatter, illustrating the quality of rendering and distribution of colors. On the other hand, the histograms are very sharp for fake cartoons showing a mismatch between fake and real rendering. To show the movement differences in these cartoons, the third image in Figure 1 shows the optical flow of the images shown on top with their 5th last frames. We can see that there is movement in real cartoons as captured by the optical flow while the only movement in fake cartoon is the change of face direction of tom.

A large set of **Explicit Fake Videos (EFV)** were downloaded from a popular Reddit post with the title “*What is up with the weird kids’ videos that have been trending on YouTube lately?*” [13]. We obtained another pool of videos by combining the names of famous cartoons with an explicit keyword such as kiss, pregnant, bikini, love story, etc. Finally, more videos were collected through YouTube recommendations and suggestions as a result of watching these initial seed videos.

In order to obtain a sample of **Violent Fake Videos (VFV)**, 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 named *Mondo Media* 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. However, it is important to note that *Happy Tree Friends* and videos from *Mondo Media* were taken from YouTube and were not found on YouTube Kids. They were used to train our model so we can classify future videos containing questionable content with greater accuracy. Table I lists the exact number of videos we collected in each category. While there is an imbalance in sizes of the categories, the proposed classifier takes segmented scenes into account (not complete videos) and the scene count is quite balanced so as to provide an unbiased classification scheme.

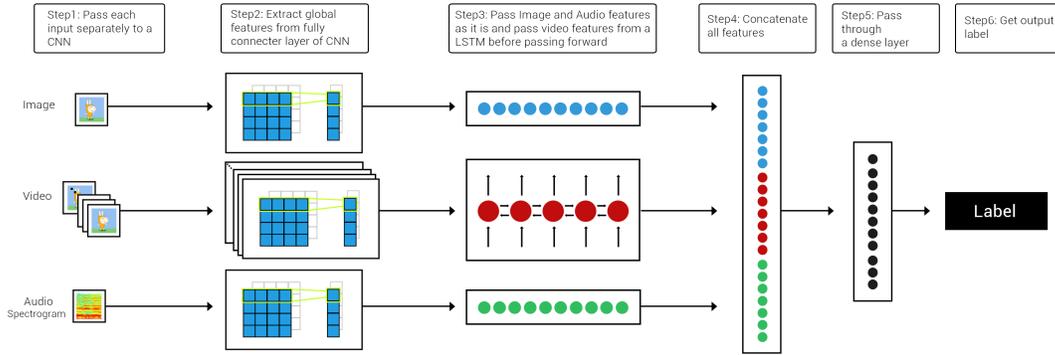


Fig. 2: Model Overview: Various layers of the model along with the steps involved at each layer can be seen. Image-based, video-based and audio spectrogram features are taken as inputs and passed through a sequence of neural networks. The features are then concatenated and a final label is assigned.

Data Set	Videos Count	Segmented Scenes
Original Videos	567	20552
Fake (All Types)	632	6529
Fake Explicit (EFV)	350	2076
Fake Violent (VFV)	150	2505

TABLE I: Count of videos in the dataset per category. Original cartoons usually had longer length per cartoon whereas fake cartoons are usually shorter. Also, there were a total of 6992 scenes from explicit videos but we removed the ones that did not have explicit content thus leaving us with a smaller number of explicit scenes.

IV. METHODOLOGY

Data Processing A video contains hundreds of frames, a large number of which are redundant i.e., movement does not change after every frame. Processing all frames through a neural network is very challenging and time-consuming. Moreover, Recurrent Neural Networks (RNNs) are plagued by the vanishing gradient problem in the presence of long sequences [14]. Hence, we split videos into *sequences of frames* called **scenes** that are used for classification.

Scene and Audio Segmentation: Each video is segmented into clips that constitute a complete scene. We use frame histograms to represent visual features of video frames and use cosine-distance between consecutive frames to detect changes in scenes. Each histogram is compared with the mean of the previous three histograms and any deviation above a threshold T marks the beginning of a new clip, as shown in Figure 3. While the exact value of the threshold T , depends on the *contrast*, *saturation* and *brightness* of a video, for our system we have empirically determined a threshold $T = 0.15$ (results excluded for brevity). We took two sets of cartoons with similar (Smurfs, Octonauts) and contrasting color palettes (Mickey Mouse, Tom and Jerry) and averaged the threshold T determined through manual analysis. However, T can be considered as an input parameter, and must be adjusted accordingly. As a simplifying assumption, we only keep scenes that contain more than a certain number of frames. For each video clip segmented, we extract its audio and convert it into a spectrogram to be used by our model.

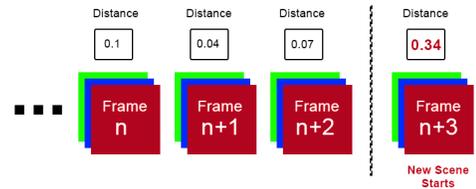


Fig. 3: Illustration of scene segmentation. Each scene comprises multiple frames. We also segment the associated audio.

Deep Learning Model: Figure 2 shows our deep learning architecture for scene classification. Inputs are passed through a pre-trained Convolutional Neural Network (CNN) for feature extraction. Frame (single image) and audio features are passed as it is to the concatenation layer. For a video clip, since we have a sequence of frames, we need additional sequence processing. For that, we pass the features extracted from the CNN to a Recurrent Neural Network (RNN) which processes the sequence and converts it into a 1D vector, which is then passed to the concatenation layer. At the concatenation layer, all features are concatenated before being passed to a fully connected layer followed by a softmax layer. Concatenation is done to help the model leverage information from different sources. Below, we explain each step in detail.

Input Feature Extraction: Deep learning models require vectors of numbers as inputs. To convert our clips, audios and images into vectors, we extract their features from a pre-trained CNN. In CNNs, filters of various sizes are convolved over an image to extract meaningful features. These features range from edge detection to global features such as shapes of objects. For a more detailed understanding of CNNs, we refer the reader to [15]. In our case, we use a pre-trained CNN to extract a feature vector for each frame in the video and for the audio spectrogram. As shown in [16], CNNs trained on a large dataset have transferable features to other domains with small amounts of training data. We use three different types of inputs in our model.

1-Frame Features: To extract features of a frame, we pass it through a pre-trained CNN. For each scene, a frame is just a randomly chosen image from the scene. Having individual frames makes it easy to recognize any fake frames, if embedded into the cartoon. In our observation, fake videos

Content	Features	Accuracy	Precision +	Precision -	Recall +	Recall -	AUC Score
Fake (All Types)	Frame	66.21	62.67	76.03	87.89	42.32	74.70
	Audio	82.04	85.83	78.65	80.15	84.68	90.17
	Movement	87.37	85.21	89.95	91.03	83.56	94.14
	Frame + Audio + Movement	92.83	94.68	90.76	91.94	93.88	97.30
Fake Explicit (EFV)	Frame	80.46	89.13	75.87	63.5	92.25	89.20
	Audio	75.85	74.32	76.92	68.22	88.74	83.68
	Movement	93.48	93.78	93.27	91.18	95.22	97.68
	Frame + Audio + Movement	90.95	92.80	90.0	86.98	94.55	96.10
Fake Violent (VFV)	Frame	81.23	84.21	86.12	88.64	81.74	86.72
	Audio	78.21	80.22	81.64	83.12	74.59	84.23
	Movement	91.54	93.70	95.00	91.22	89.67	95.76
	Frame + Audio + Movement	95.90	95.26	96.79	96.21	96.89	97.54

TABLE II: Summary table showing the performance of our deep learning model on the **test** data set. There are three types of content where we perform classification. Table does not show Benign Fake Videos (BFV), as any video that is fake but not classified as violent or explicit is automatically assumed to be a BFV.

are rendered differently than normal ones. Using this insight, we are able to clearly distinguish between them based on the frame features.

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. Moreover, one needs to look at a sequence of frames to determine any explicit/violent activity. Movement features are extracted after each frame is passed through the CNN and saved in a 2-dimensional array of size $n \times d$ where n is the number of frames in a scene and d is the dimension of output feature vector by the CNN. This sequence of feature vectors is passed through a RNN to get a global representation of the movement.

3-Audio Features: To extract audio features, each frame’s audio spectrogram is passed through the CNN and its output is 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. Hence, audio features are an important factor in distinguishing fake and inappropriate videos.

Implementation Details: For the choice of pre-trained CNN, we have used *VGG19* [17], which has been trained on the ImageNet. The other choices were VGG16, Resnets, and Google LeNet but we empirically found out that VGG-19 worked best for our problem. For feature extraction, we use the “fc2” layer of the CNN. For RNNs, we use bidirectional Long Short Term Memory Networks (LSTMs) with 256 units. The final dense layer has 512 units. The code has been written in *Keras* library [18]. We plan on open sourcing the code and the dataset soon. The training is run for 25 epochs with a learning rate of 0.0001. We have used the *Adam* optimizer [19]. Our model predicts classes in a hierarchical way. First, it categorizes a video into real or fake. Next, for fake videos, it classifies them into explicit, violent, or appropriate (benign). We split our dataset into 70% training, 20% testing and 10% validation set and performed 5-fold cross validation. After every layer, we have used a dropout of 0.5. We also balance the number of scenes from each category during training to prevent classification bias.

V. RESULTS

Table II shows the summary of our results. Fake Benign (BFV) is not included because if a video is labeled fake but not explicit or violent, then it is considered a benign fake video. For each type of content in our data set, we trained our neural network for 1 vs. all classification and as evident in the table, the results are consistently above 90% for all the categories and all the metrics used. In fact, at times a single frame is enough to enable the model to classify correctly. This happens when real and fake frames have pronounced differences, as shown in Figure 1. For example, in a scene where two characters are kissing each other, the scene consists of the characters first coming close to each other followed by a kiss. The part where the characters are kissing each other is explicit. Therefore, if the random frame is chosen from the set of frames where there is kissing, the model will perform well. Similarly, there are some cartoons whose fake version has very different rendering compared to the original version. In such cases, the model easily learns any distinguishing patterns based solely on a single frame. Looking at the results for audio features, they seem to have a better average discriminatory potential for content classification than images. In the case of fake content, the accuracy is considerably greater than frames. It remains decent on both explicit and violent videos. Results using movement features are consistently good in all scenarios, which is expected since every scene has some movement that indicates the nature of the content present in it. For example, in violent videos, a large number of scenes have blood flowing or body parts being cut. In explicit videos, many scenes contain kissing, touching, laying on a person, etc. All of these scenes have movements that should be leveraged for classification and our neural network is able to use these movements. Finally, we can see that the combined features always outperform the individual features except in the case of explicit videos. This happens because the neural network is able to use all three types of features and figures out which features are important for a particular scene and which are not. However, in explicit videos, we hypothesize that the audio is adding noise to the model thus causing reduction in accuracy albeit the reduction is very small. Hence, this proves the point that multiple video features are necessary to successfully



Fig. 4: Class Activation Map capturing Minnie Mouse’s provocative clothing and some other explicit scenes.

attribute videos to different categories. Moreover, this makes the detection method robust against changes in different types of input features as we will show in the next section.

A. What does our neural network see?

In vision-based ML problems, a common issue is that the model at times, learns patterns that are orthogonal to the actual patterns observed by a human being to detect an inappropriate video. For example, if a network has learned a pattern related to the background color of an explicit video, it is highly irrelevant because (background color is mostly inconsequential). Thus, it is desirable that the neural networks focus on the most appropriate and relevant regions in a frame. To this end, we use Class Activation Maps (CAMs) [20], which highlight the regions of an image that are of relevance to a particular class. In other words, a CAM accentuates those regions of a frame that the CNN focused on to strengthen its classification decision. To confirm that our classifier was indeed focusing on the right regions, we generate the CAMs of different frames in a scene involving Minnie Mouse. As evident in Figure 4, our classifier is using the most relevant regions for labeling the images. The explicit nature of a scene is clearly captured and actually used during classification.

VI. RELATED WORK

An important line of work are standard porn and gore detectors. However, mapping those approaches to our purpose would lead to a popular problem in Computer Vision known as domain mismatching [21] in which a model is trained on one domain and tested/used on a different domain. The challenge is that videos from both the domains are not the same. Therefore, a classifier trained on a different set than where it will be used, does not perform well. Also, a standard way of detecting pornographic content has been skin detection [22]. However, such a solution fails to recognize that people may show a lot of skin in activities that have nothing to do with sex (e.g., sunbathing, swimming, running, wrestling), leading to a lot of false positives. Conversely, some sexual practices involve very little exposed skin, leading to unacceptable false negatives.

Our work is closest to [5], where two approaches are presented to detect unsafe content on YouTube. One based on supervised classification using video-level, user-level and comment-level features and second, based on Convolutional Neural Network using video frames.

VII. CONCLUSION

YouTube Kids (YTK) has evolved as the go to app that kids use during their free time. However, it has been observed that sometimes content that is deemed unsuitable for children

can land on YTK and other similar apps. Hence, a more comprehensive filtering mechanism needs to be devised. To this end, we first explore the tricks and tactics used by content producers by collecting a dataset of unsuitable videos on YTK. Based on our findings, we present a deep learning based approach that can flag unwanted content. The approach is robust and performs well despite limited training. We believe that our system can serve as a good first step for filtering content and alerting for manual review.

VIII. ACKNOWLEDGMENTS

This research was supported in part by NSF grant IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

- [1] “Familyzone - kids under 8yo spend 65 percent of their online time on youtube,” <https://tinyurl.com/y7pivzep>.
- [2] “The newyork times - youtube kids, criticized for content, introduces new parental controls,” <https://tinyurl.com/yarlpjja>.
- [3] K. Orphanides, “Wired - children’s youtube is still churning out blood, suicide and cannibalism,” March 2018, <https://tinyurl.com/y7jsm4vs>.
- [4] “Techcrunch - youtube kids app reported to fit for featuring videos with adult content,” <https://tinyurl.com/yccpmwoc>.
- [5] R. Kaushal *et al.*, “Kidstube: Detection, characterization and analysis of child unsafe content & promoters on youtube,” *CoRR*, vol. abs/1608.05966, 2016.
- [6] “The week - does elsgate prove youtue is too big to control,” <https://tinyurl.com/y9vw65m2>.
- [7] “Boston25 news - deceptive online videos with popular kids characters hide sinister scenes,” <https://tinyurl.com/y6w79nq4>.
- [8] “Polygon - youtube’s big problem has always been its algorithm, not kids programming,” <https://tinyurl.com/ybwupeqj>.
- [9] “Youtube tops 1 billion hours of video a day, on pace to eclipse tv,” <https://tinyurl.com/j2bvvyo>.
- [10] “The verge - inside elsgate, the conspiracy-fueled war on creepy youtube kids videos,” <https://tinyurl.com/ya54w72b>.
- [11] “Medium - something is wrong on the internet,” <https://tinyurl.com/ya6qwtww>.
- [12] “The NewYork Times -The Flourishing Business of Fake YouTube Views,” <https://tinyurl.com/y34v8mjq>.
- [13] “Reddit - loop,” <https://tinyurl.com/yafktawn>.
- [14] Y. Bengio *et al.*, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, March 1994.
- [15] Y. LeCun *et al.*, “The handbook of brain theory and neural networks,” 1998, pp. 255–258.
- [16] M. Oquab *et al.*, “Learning and transferring mid-level image representations using convolutional neural networks,” in *CVPR*, June 2014.
- [17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale visual recognition,” *ICLR*, 2015.
- [18] “Keras Documentation,” <https://keras.io/>.
- [19] D. P. Kingma *et al.*, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.
- [20] B. Zhou *et al.*, “Learning deep features for discriminative localization,” in *CVPR*, 2016.
- [21] V. Jain *et al.*, “Online domain adaptation of a pre-trained cascade of classifiers,” in *CVPR*, 2011, pp. 577–584.
- [22] M. Perez *et al.*, “Video pornography detection through deep learning techniques and motion information,” *Neurocomputing*, vol. 230, 2017.