ENGINEERING RESEARCH JOURNAL (ERJ)

InspectorNet: Transformer network for violence detection in animated cartoon

Mahmoud M. Taha [a], Abdulwahab K. Al-Sammak [a] and Ahmed B. Zaky [b]

[a] *Faculty of Engineering at Shoubra, Cairo, Egypt*
[b] *Egypt japan university of science and technology computer science and information technology programs (CSIT) assistant professor on leave from the Shoubra faculty of engineering*

Abstract. InspectorNet is a convolutional neural network based on transformer deep learning techniques, which is designed to address some of the limitations of current state-of-the-art artificial neural networks (ANN) models. The paper compares the performance of InspectorNet against a commonly used neural network in image classification, ResNet [1], on the Danbooru2020 dataset of animated cartoon images with a variable number of classes. The comparison shows that while both networks require significant computing resources for training, InspectorNet demonstrates better classification performance in certain test situations. The paper also highlights that with the increasing access to the internet, it is important to control the dissemination of sensitive content such as violence, but current neural networks may not be as effective in filtering cartoon movies aimed at children as the filters for these movies are different from those for adult movies. InspectorNet also has a compact architecture than many modern networks, such as ResNet, which results in better performance on low-resource devices.

Keywords: violence filtering, transformer, image classification.

1. INTRODUCTION

Distributing inappropriate video footage, even with consent, is illegal and dangerous. It can be especially harmful to minors if it involves violence or child sexual abuse. Law Enforcement Agencies (LEAs) are required to thoroughly review evidence before bringing a criminal case when there is a possibility that illicit content was created or disseminated. Violence is defined as any commercial product, such as a fictional drama, that promotes violent behaviour. Studies such as paper [2] have shown that Generation Z, which includes individuals between the ages of 11 and 22, is particularly at risk due to their heavy use of the internet.

Efforts to filter and control explicit content have had limited success due to its prevalence among adults and children. With the advancement of computer vision and deep learning, identifying and filtering unwanted content has become a key objective in this field of study. Identifying violence in movies can be challenging due to the variety of content and variations in quality. Supervised classification is a significant challenge in machine learning, as it may lead to false-positive or false-negative results in certain situations, such as wrestling. Traditional methods used for video filtering like paper [3] only work on one-dimensional attributes, which can be ineffective in some cases.

The InspectorNet network is unique in that it prioritizes the classification of animated cartoon images and movies, and can operate on low-resource devices such as mobile phones and gaming consoles. To achieve this, it has a lightweight and precise design. This is important as children may find it difficult to categorize animated cartoon content as threatening, thus watching violent movies may be disturbing for them. This classification task serves as a foundation for other computer vision problems such as detection, localization, and segmentation.

In this study, we develop InspectorNet, a deep-learning neural network that classifies violent scenes automatically in order to filter violent content in animated movies. As far as we are aware, this is the first attempt to use convolutional neural networks to categorise animated cartoons aimed at children. Our contribution involves classifying violent animated cartoons using convolutional networks dubbed InspectorNet. Five sections make up the remainder of this study. Section 2 discusses a review of earlier studies on the issue of violence detection, and Section 3 gives an overview of the most popular datasets for the categorization of violence detection. Then, in Section 4, which comes next, the main components of our model InspectorNet are presented. In Section 5, we discuss the experimental setup of the suggested approach. We compare the outcomes of our network training to one of the earlier neural networks in Section 6. Finally, in section 7, we demonstrate how our suggested model might be improved in the future to improve efficiency.

2. LITERATURE REVIEW

Images are the most common medium for delivering violent content, including texts and audio. As violence often includes skin exposure, skin detection is often used to identify perpetrators. However, skin detection is challenging due to the limited generalizability of the methods used. In the study by Vu Lam and Duy-Dinh Le, MediaEval [4] was presented which combined trajectory-based motion features with SIFT-based and audio characteristics. Their findings suggest that motion parameters based on trajectory are still relatively effective. Combining image and audio data can improve overall performance for detecting violent events in videos.

To identify illegal content in a perpetrator's documents, Law Enforcement Agencies (LEAs) use hashing algorithms. These techniques are computationally efficient, resistant to producing false positive alarms, and can only identify content

that is similar to what is in the LEA's database. Due to the sensitive nature of not-safe-for-work (NSFW) content, large datasets are gathered in close cooperation with LEAs. As a result, data-driven algorithms often produce false positives or false negatives when detecting NSFW. Convolutional neural networks (CNNs) are widely used in image classification and have a high accuracy rate, but they also have limitations, such as the accumulation of characteristics at each successive layer [5].

***TABLE 1**. Inappropriate content filtering list of previously used methods*

Paper	Target	Result
Shen et al. [7]	Detection of violence using SSD extracted features	94.7% accuracy on NPDI dataset
Moustafa et al. [8]	Adult Video detection using CNN models AlexNet & GoogleNet	94.2 % on NPDI-800 dataset
Kejun et al. [9]	Detection of explicit Female breast using Classical hand-crafted features	FPR: 7.46% and FNR: 4.86%
Mahadeokar et al., 2016 [10]	Violence detection using Thin ResNet-50	FPR: 12.8% and FNR: 5.53% using Violence-2k
Mallmann et al., 2020 [11]	Private Organs detection using Faster R-CNN Inception v2	FPR: 1.66% and FNR: 2.34% using Violence-2k

Previous work in this field can be divided into two categories. The first is based on the type of content that needs to be filtered, such as guns, blood, fights, or all of them. The second group focuses on specific media forms, like photographic videos, hand-drawn graphics, or animated cartoons. According to Peter [6], there is a need to pay closer attention to concerns about identification and teenagers' exposure to

sexually explicit online content or violent movies. Increased exposure to violent scenes or sexually explicit internet content is linked to more violent behaviour. A standard set of explicit or violent object classes and a related benchmark dataset have been defined, as shown in Table 1, which lists the objectives and results achieved by the studied approaches.

There are still many challenges in the video filtering industry, as there are a wide variety of video formats. Identifying what exactly will be filtered in movies is a crucial question to answer, as there are a number of alternatives such as guns, weapons, alcohol, blood, injuries, fighting, shouting, action detection, violent action recognition, or simply detecting everything. Additionally, there are several subcategories such as suggestive violence or explicit violence. As seen in Fig 1, there are various types of cartoon movies and animated graphics, including 2D films, 3D videos, stop motion, and line drawings. An animated picture may be considered as a distorted version of a real image. The main goal of our experiment is to evaluate InspectorNet's performance on animated-cartoon images in DeepDanbooru dataset [16], and we will compare it to previously used methods in video filtering field.

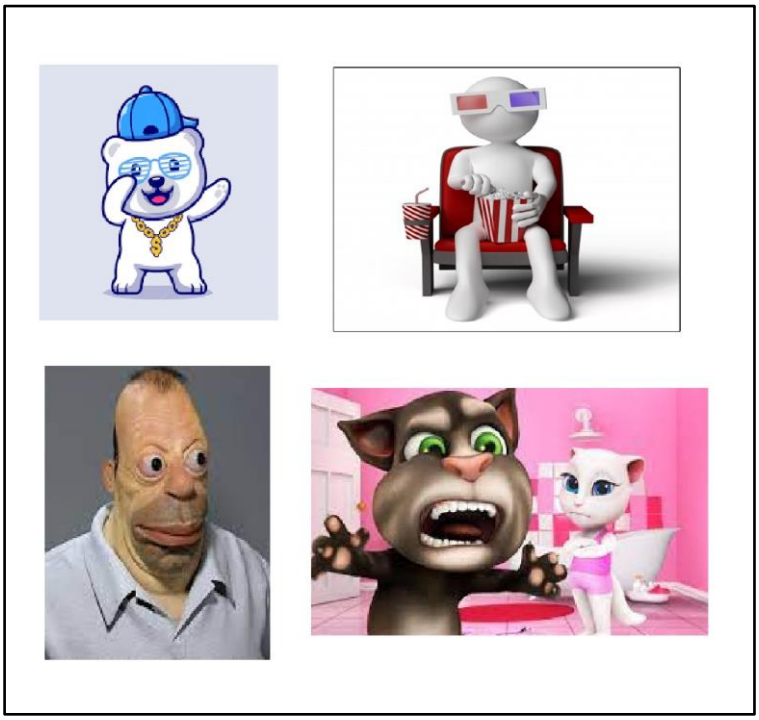


FIGURE 1*. Examples of drawing styles and animated images.*

A Transformer network is a state-of-art architecture that solves sequence-to-sequence tasks while handling long-range dependencies with ease. [13] [14]. As shown in Fig 2, The encoder-decoder design of the majority of competing neural sequence transduction models is akin to the transformer network architecture. The encoder converts the input sequence of symbol representations into a series of continuous representations in this case. Several symbols are then generated by the decoder. For both the encoder and decoder, the Transformer employs layers that are point-wise completely coupled. Two attention network kinds exist in the transformer network. Scaled Dot-Product Attention and Multi-Head Attention, consist of several parallel attention layers. We produce the transformer attention function's output matrix in the manner suggested by equation 1.

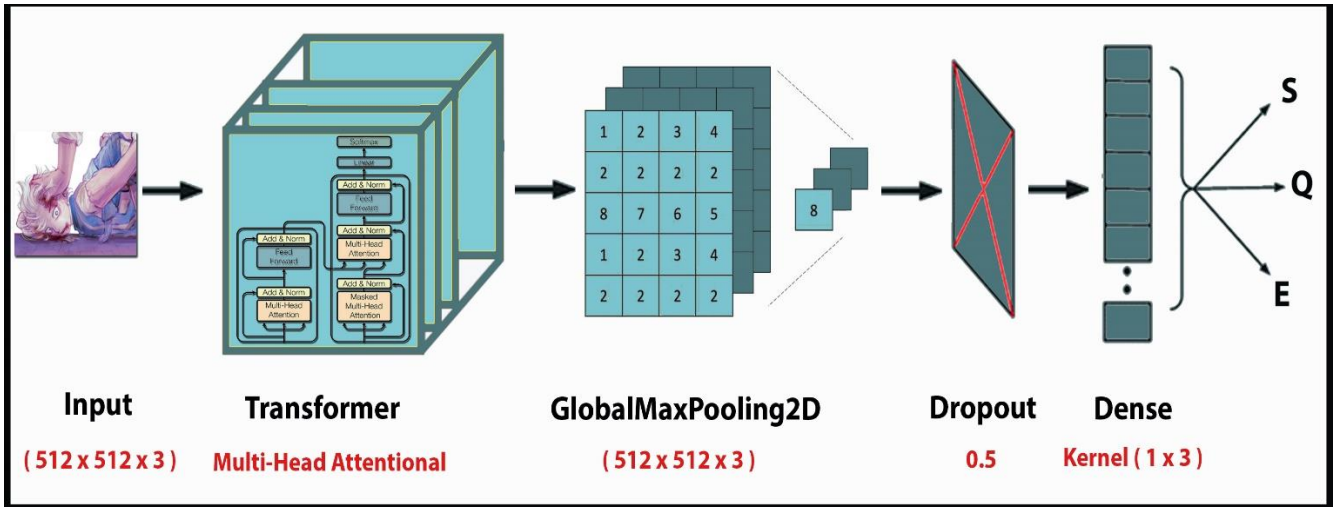


FIGURE 2. transformer model architecture used in InspectorNet

EQUATION 1. Transformer attention function output equation

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

3. DATASET

Obtaining a reliable animated cartoon dataset may be difficult as most datasets in the scientific field are for real people films [15]. As seen in Table 2, deep learning in computer vision requires large annotated datasets. Classification and categorization have benefited from the creation of ImageNet, but a similar dataset with violent tags or labels for a large collection of photographs is not available, preventing the learning and detection of more information about images. Categorization and classification are rough descriptions of an image. The most well-known dataset that meets these characteristics is Danbooru [16]. The Danbooru dataset is used for different classification tasks in animated cartoons field including violent action classification. For violence detection, each image in Danbooru [16] has one of three categories: safe images (like Fig. 3), which don't contain any

violence or nudity, questionable images (like Fig. 4), which may or may not have concealed violence, and explicit images (like Fig. 5), which contain explicit violence.

***TABLE 2**. List of datasets used in inappropriate content analysis [15]*

Name	Datatype	Description
Danbooru2018 [16]	images	A 300 GB Crowdsourced and Tagged 3.33m+ anime images illustration dataset classified into 3 categories: safe, questionable and explicit (violent).
VSD2014 [17]	videos	A dataset of violent videos or most popular Hollywood films.
kinetics [18]	videos	A dataset of various types of all human action recognition videos.
UCF-101 [19]	videos	A big dataset of various types of human action recognition videos.
youtube8m [20]	videos	A big dataset of various types and activities tagged from YouTube.

***FIGURE 3**. Safe Image from Danbooru dataset*

FIGURE 4. Questionable Image containing gun from Danbooru dataset

FIGURE 5. Violent action Image from Danbooru dataset

4. INSPECTORNET TASK AND ARCHITECTURE

In order to assist parents and law enforcement agencies in automating the process of identifying and filtering violent content in animated cartoons, we aim to develop a fast and accurate solution for recognizing violent actions and weapons. Our proposed method will also specify the level of violence in the picture content by using a transformer network based on a softmax neural network. Additionally, we will demonstrate how the acquired characteristics of this model can be applied to other tasks, such as photo classification. InspectorNet is mainly dependent on the

transformer network so it can better retain and simulate hierarchical relationships within the internal knowledge representation of a neural network.

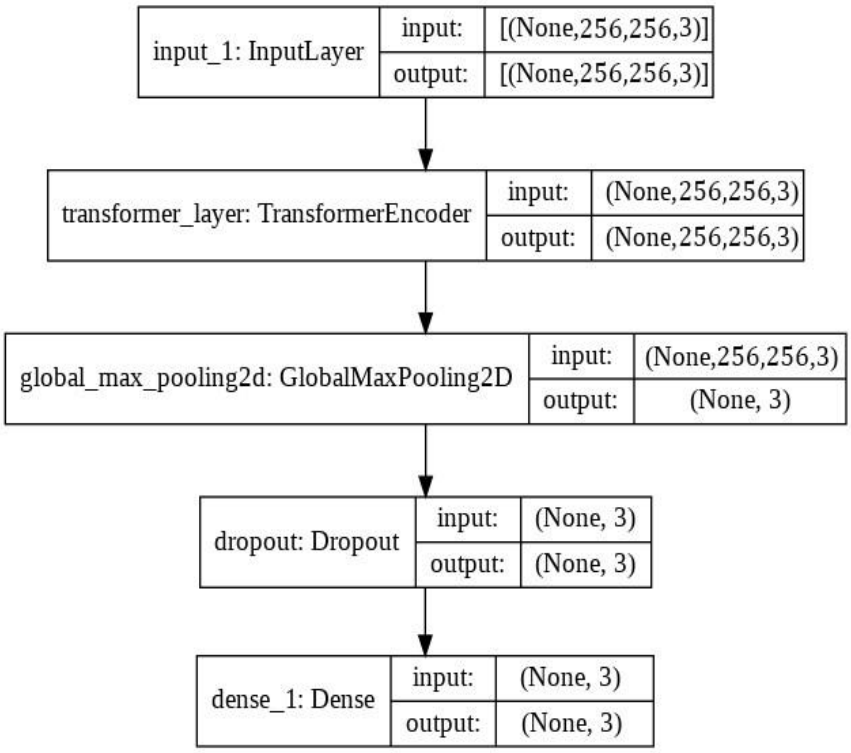


FIGURE 6*. InspectorNet architecture based on transformer network*

As depicted in Fig 6, InspectorNet is composed of a transformer network combined with max pooling and dense neural layers. The transformer network is responsible for identifying image attributes and classifying individual frames based on previous frames. The dense layer works in conjunction with the transformer network to determine the correct classification for each image, similar to a traditional CNN. Global average pooling is a more natural fit for the convolutional structure than fully connected layers, as it requires correspondences between categories and feature maps. The feature maps are therefore just subcategories of confidence maps. The dropout layer serves as a mask, preserving all other neurons but removing some

neurons' contributions to the following layer. The output shape will depend on the number of trained neurons or units in the Dense layer.

5. IMPLEMENTATION

5.1 *Software and hardware*

To run the classification algorithms, we used a Dell laptop and virtual instances hosted on the Google Colab Platform. The Nvidia GeForce 1050 Ti Mobile GPU, 16GB of RAM, and Intel Core i5 Extreme 7th generation CPU are all included in the Dell laptop. Instances in Google Colab may be set up with a single Nvidia Tesla K80, 5GB of RAM, and an Intel Xeon (2.0 GHz) CPU. The primary programming environment we used was Colab IDE, while the programming packages we utilised were Python 3.6, Keras 2.1.5, and Tensorflow 2.5.

5.2 *Preprocessing and Training architecture*

Based on the dataset's characteristics and the requirements of the InspectorNet algorithm, a variety of preparation techniques were applied. Table 3 summarizes the preprocessing techniques applied to each dataset. For training, validation, and testing, dataset is divided into 65%, 15 % and 20 %. Adam method [21] was used as a cutting-edge optimizer.

TABLE 3*. Preprocessing applied to dataset*

preprocessing	image samples
Min-max norm [0,1]	Applied
Grayscale	Images are fed to InspectorNet in RGB mode
Resize	resized to 256 x 256

6. RESULTS AND EVALUATION

The Danbooru dataset, which has been regularly used to evaluate illegal picture content detection systems, was used to test InspectorNet. Classification accuracy was used as a benchmark metric. The complete classification results are displayed in Table 4. To further clarify the results and display the true positive and true negative data, we also created Figure 7 and Figure 8 to show confusion matrix of InspectorNet and ResNet respectively. According to Table 4, we can more accurately categorize violence using the InspectorNet architecture than ResNet for all common categories like guns or violent actions. However, InspectorNet requires more time and computer resources to achieve its objectives. InspectorNet can deliver more accurate classification when the image resolution is increased to 1024 pixels. Ultimately, InspectorNet accuracy was 97.2%, which is higher than ResNet.

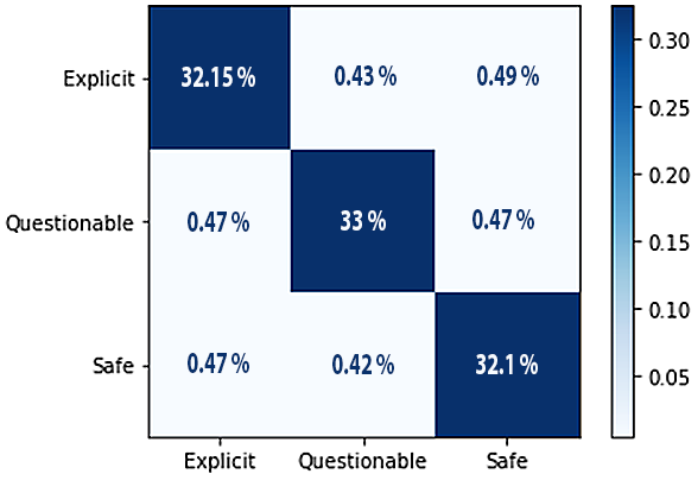
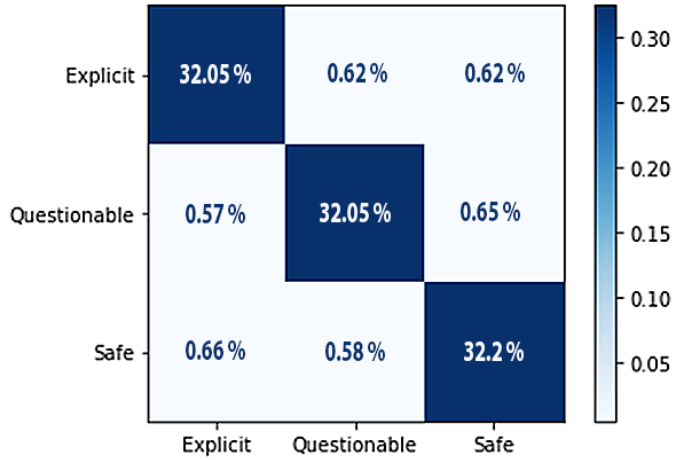


FIGURE 7*. InspectorNet violence classification confusion matrix*



FIURE 8. *ResNet violence classification confusion matrix*

TABLE 4*. InspectorNet and ResNet results on Danbooru*

	InspectorNet	ResNet
Classes	3	3
Instances	80000	80000
Training period	120 hours	80 hours
ACC (%)	97.2	96.3

7. CONCLUSION AND DISCUSSION OF RESULTS

Our results indicate that InspectorNet can effectively classify violent and inappropriate situations even with a small number of hyper-parameters. Figure [9] shows that InspectorNet shows better performance compared to many neural networks on different datasets. InspectorNet's value is that it has a compact design that can function well with better accuracy on low-resource devices such as smartphones in contrast to many current neural networks like ResNet. Small variations among classes could also benefit both the baseline and InspectorNet designs for animated-cartoon datasets. In terms of training times, we estimate that

using a Tesla K80 GPU, it should only take a few days to achieve good accuracy on three classes with 100k pictures each (Danbooru dataset). The variety of production methods used to create cartoon films makes this task challenging for most neural networks, which is what allows InspectorNet to outperform other neural networks such as ResNet.

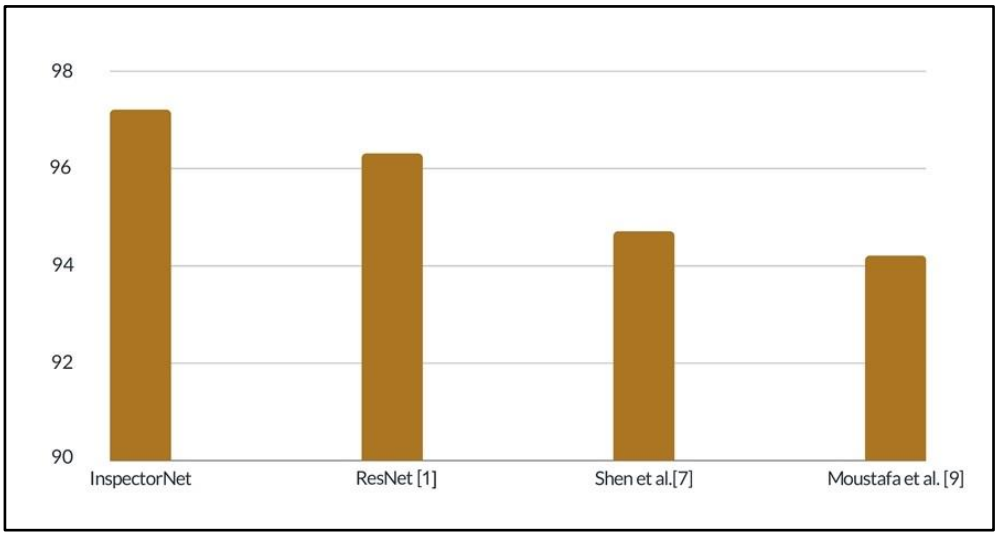


FIGURE 9. *accuracy comparison between results of InspectorNet on Danbooru, ResNet [1] on Danbooru, Shen et al. [7] on NPDI dataset and Moustafa et al. [8] on NPDI-800 dataset*

8. CHALLENGES

To improve classification accuracy, deep learning and algorithms for blood and guns in violence can be combined. This will help to eliminate incorrect classifications for certain behaviors, such as red juices or toys. However, the deep learning method that only uses static data remains one of the most competitive action recognition techniques. Future research in this field should include various elements in the classification process, such as speech recognition, video subtitles, and voice tone classification. There are many types of animated images such as stop-motion and line drawings, as shown in Fig 1.

Our research has shown that when it comes to classifying violence, InspectorNet can still outperform other traditional machine learning techniques as well as neural networks that have been specifically designed and optimized. However, running the training session over photographs at their original size becomes very resource-intensive as the dimensions of input images increase, due to the exponential increase in resources required.

REFERENCES

1. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
2. Nor Azura Ab Rahman, Tengku Maaidah Tengku A Razak, Mohd Shahrudin Mohmud, Nur Ulfah Harun, Abi Yazid Tukiran, Nurhidayah Muhammad Hashim, and Rosmawati Mohamad Rasit. The implications of pornography addiction among adolescents. Journal of Positive School Psychology, 6(3):8904–8913, 2022.
3. Kamrun Nahar Tofa, Farhana Ahmed, Arif Shakil, et al. Inappropriate scene detection in a video stream. PhD thesis, BRAC University, 2017.
4. Claire-Hel´ ene Demarty, C` edric Penet, Markus Schedl, Ionescu Bogdan,´ Vu Lam Quang, and Yu-Gang Jiang. The mediaeval 2013 affect task: violent scenes detection. In MediaEval 2013 Working Notes, page 2, 2013.
5. Rinat Mukhometzianov and Juan Carrillo. Capsnet comparative performance evaluation for image classification. arXiv preprint arXiv:1805.11195, 2018.
6. Jochen Peter and Patti M Valkenburg. Adolescents' exposure to sexually explicit internet material, sexual uncertainty, and attitudes toward uncommitted

sexual exploration: Is there a link? Communication Research, 35(5):579–601, 2008.

7. Rongbo Shen, Fuhao Zou, Jingkuan Song, Kezhou Yan, and Ke Zhou. Efui: An ensemble framework using uncertain inference for pornographic image recognition. Neurocomputing, 322:166–176, 2018.
8. Mohamed Moustafa. Applying deep learning to classify pornographic images and videos. arXiv preprint arXiv:1511.08899, 2015.
9. Xin Kejun, Wu Jian, Ni Pengyu, and Huang Jie. Automatic nipple detection using cascaded adaboost classifier. In 2012 Fifth International Symposium on Computational Intelligence and Design, volume 2, pages 427–432. IEEE, 2012.
10. Jay Mahadeokar and Gerry Pesavento. Open sourcing a deep learning solution for detecting nsfw images. Retrieved August, 24:2018, 2016.
11. Jackson Mallmann, Altair Olivo Santin, Eduardo Kugler Viegas, Roger Robson dos Santos, and Jhonatan Geremias. Ppcensor: Architecture for real-time pornography detection in video streaming. Future Generation Computer Systems, 112:945–955, 2020.
12. Deepdanbooru-v3, 2022. https://github.com/KichangKim/DeepDanbooru/releases.
13. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
14. Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3163–3172, 2021.
15. Abdel Wahab Alsammak Mahmoud Mohammed Taha, Dr. Ahmed B. Zaky. Filtering of inappropriate video content a survey. INTERNATIONAL

JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT), 11(02), 2022.

16. Danbooru2018, 2018. https://www.gwern.net/Danbooru2018.
17. Vsd2014 dataset, 2014. https://www.technicolor.com/dream/researchinnovation/violent-scenes-dataset.
18. kinetics dataset. https://deepmind.com/research/open-source/opensource-datasets/kinetics/.
19. Ucf-101 dataset. https://www.crcv.ucf.edu/research/data-sets/humanactions/ucf101/.
20. youtube8m dataset. https://research.google.com/youtube8m/.
21. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.