International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Advancements in NSFW Content Detection: A Comprehensive Review of ResNet-50 Based Approaches

Sanjay A. Agrawal[1], Vaibhav D. Rewaskar[2], Rucha A. Agrawal[3], Swapnil S. Chaudhari[4], Yogendra Patil[5], Nidhee S. Agrawal[6]

Submitted: 08/05/2023 **Revised**: 15/07/2023 **Accepted**: 07/08/2023

Abstract: The exponential growth of explicit images posted on social media is increasing day-by- day. With children and minors having unrestricted access to the internet, the rapidly rising availability of pornographic content has created many difficulties in modern life. Therefore, there is a need to build a system that will detect the explicit content in an image and the text present in the image. In this system a deep learning-based architecture is used to detect the images and text in images. The proposed system employs a pre-trained convolutional neural network (CNN) model known as ResNet50 to classify the image as safe or not safe. The existing system used only CNN in image detection, and it resulted in less accuracy, whereas our proposed system uses ResNet50 and is expected to give more accuracy as compared to the existing system.

Keywords: *ML: Machine Learning, CNN: Convolutional Neural Network, ResNet: Residual Network, SFW: Safe for Work, NSFW: Not Safe for Work*

1. Introduction

Modern technology has greatly enhanced the availability of information and its accessibility online. Additionally, it is concerning that people now have better access to both general information and adult information, particularly for young people. Developing a system capable of detecting and filtering out adult content from a vast amount of data is one solution to address the challenge of preventing children from accessing inappropriate material. The widespread use of digital devices and the internet has made it easier for children to access adult content, which can have adverse effects. The objective of this system is to create a classifier that can determine if an input image is "Safe for work" or "not Safe for work," as well as to categorize the improper photos based on their offensive content. In order to categorize photos as safe or not safe, the current system uses CNN. However, this current approach has the issue of vanishing gradient, which causes some crucial aspects of images to be lost. We use the ResNet-50 algorithm, which provides better accuracy than the CNN model, to solve this issue. This model gives the probability of the image being safe or not

[1] *Assistant Professor, Marathwada Mitra Mandal's Institute of Technology, Pune, Maharashtra, India*
ORCID ID : 0000-0002-7499-8702
[2] *Assistant Professor, Marathwada Mitra Mandal's Institute of Technology, Pune, Maharashtra, India*
ORCID ID : 0009-0002-5436-2958
[3] *Assistant Professor, Marathwada Mitra Mandal's Institute of Technology, Pune, Maharashtra, India*
ORCID ID : 0009-0008-2721-913X
[4] *Assistant Professor, Marathwada Mitra Mandal's Institute of Technology, Pune, Maharashtra, India*
ORCID ID : 0000-0002-5636-7384
[5] *Assistant Professor, Marathwada Mitra Mandal's Institute of Technology, Pune, Maharashtra, India*
ORCID ID : 0000-0001-8735-591X
[6] *Assistant Professor, G H Raisoni Engineering & Management, Wagholi, Pune, Maharashtra, India*
ORCID ID : 0009-0009-7082-5634
** Corresponding Author Email: sanjay.agrawal@mmit.edu.in*

2. Literature Survey

2.1. A Deep Learning Approach to Identify Not Suitable for Work Images.

Although some images may include nudity and pornography, which can be considered unsuitable for work (NSFW) and possibly offensive to viewers, this research introduces a method for categorizing such NSFW images found on Arquivo.pt using deep neural network techniques. By utilizing Arquivo.pt statistics, a large image dataset was compiled, and pre-trained neural network models, such as ResNet and SqueezeNet, were evaluated and recommended for the NSFW classification task. The accuracy rates of these models were initially determined to be 93% and 72%, respectively. However, after undergoing a rigorous tuning process, the accuracy rates improved to 94% and 89%, respectively. [1]

2.2. Building a NSFW Classifier: -

It is suggested to use an explicit content detection (ECD) tool to identify media that is not appropriate for work (NSFW), such as images and videos. The suggested ECD machine is built entirely on a residual network (also known as a deep learning model), which provides the opportunity to denote the explicitness in media content. To decide whether the content is explicit or non-express, the price is also compared to a defined threshold. The testing results

demonstrate that the proposed version demonstrates an accuracy of roughly 95% when tested on our photo and video datasets. [2]

2.3. Technology for Detecting Explicit Content: A Step Towards a Secure and Moral Environment :-

An explicit content detection (ECD) system (NSFW) is recommended for identifying picture and video content that is not suitable for work. This system utilizes a residual network, also known as a deep learning model, as its foundation to provide a probability indicating the level of explicitness in the media content. The probability is then compared to a predetermined threshold to determine whether the content is explicit or not. The proposed method not only distinguishes between explicit and non-explicit content but also categorizes the level of explicitness as high, medium, or low for each media content. Additionally, the system is designed to flag media files with altered extensions as suspicious. Based on experimental results using picture and video datasets, the proposed model demonstrated an accuracy rate of approximately 95% [3]

3. Methods

3.1. Convolutional Neural Networks

We examined several CNN implantations, beginning with the most basic and progressing to the most sophisticated and, for the most part, the most accurate. A convolutional layer with the shape 4x4x3x8, a ReLu activation, a max pooling layer with the window size 8x8, stride 8, a convolutional layer with the shape 2x2x8x16, a ReLu activation, another max pooling layer with the window size 4x4, stride 4, and finally a fully connected layer with either 1 or 4 outputs made up our initial CNN implementation. The fully connected layer uses a sigmoid activation to output a classification as either NSFW or SFW when using one output.

3.2. Residual Neural Networks

The next step was to expand the network significantly in order to achieve higher accuracy. The network may learn difficult real-world data more effectively with a larger model. If the model is too big, though, there's a chance that the network will be overfit to the training data and produce more variation. The vanishing gradient problem is another issue that arises with growing models. The gradient shrinks exponentially as more layers are added, which causes the network's learning to stall. It makes intuitive sense that as layers are added, data flow via the network slows and finally gets saturated or clogged. Using a residual network is now the common fix for this issue. Between layers, a residual network adds "skip links." A skip connection, as the name suggests, adds output from an earlier layer to the input of a later layer by bypassing often two to three levels at once. Residual networks successfully address the vanishing gradient problem by facilitating data propagation from earlier layers to later layers.

3.3. Data Set Description: -

The dataset used in this study was downloaded from the Kaggle website. Kaggle essentially offers a variety of datasets that are open source and available in a number of different data formats. Data is supported across all platforms, as well. The dataset has been used in this study to develop a machine learning algorithm-based system for the not safe for work image detection. The dataset contains the various types of image folders such as neutral, porn, sexy, drawing, hentai. The neutral folder contains all different types of images of animals, humans, electronic objects etc. The porn, sexy and hentai contains vulgar images and explicit content. The drawing folder contains the handmade drawings and also the cartoon images. Those images show some explicit and non-explicit content.

3.4. Methods

The identification of photos that are unsafe for work is created using several algorithms in the current systems, but it has significant drawbacks. Utilizing methods of machine learning. Any machine learning algorithm will adhere to the following principles:-

1. Preprocessing: We utilize a database or dataset that has large-scale photos. Large photos cannot be processed, thus they must be reduced to 224*224*3 dimensions.
2. Data in the database is split into two categories, namely training sets and testing sets. Data from the training set make up 80%, while data from the testing set make up 20%.

The goal of the research is to develop a more accurate model to detect explicit and non-explicit content.

4. Proposed System

We wanted to create a larger network for better accuracy. The larger the model, the more effectively the network can learn complex real-world data. However, making the model too large can cause the network to overfit the training set, resulting in high variance. Another problem that arises in larger and larger models is the vanishing gradient problem. As layers are added, the gradient decreases exponentially and network learning stops. Intuitively, the flow of data through the network slows down with each layer added, eventually saturating/congesting the network as more layers are added. The standard solution to this problem today is to use residual networks. Residual networks add "hop connections" between layers. Figure 1.1 indicates the

learning model's recommended architecture, where a variety of machine learning methods are employed to find photos that shouldn't be utilized at work. In this model the not safe for work images dataset is considered from the Kaggle website as an input data and process the data.

Firstly, we collect the dataset from the Kaggle website, this dataset contains different types of images which are explicit and non-explicit. It includes human images, objects, drawing etc. This dataset may contain the unnecessary data or the images which are not suitable for our model to be processed. So make it suitable for our model, the data set is to be preprocessed. The preprocessing includes the resizing of images and converting it into NumPy array. These photos are transformed into a height, width, and channel formatted NumPy array. Firstly, we collect the dataset from the kaggle website, this dataset contains different types of images which are explicit and non-explicit. It includes human images, objects, drawing etc. This dataset may contain the unnecessary data or the images which are not suitable for our model to be processed. So make it suitable for our model, the dataset is to be preprocessed. The preprocessing includes the resizing after preprocessing the features are extracted from the images. These features include body parts, shape of body, color, size etc. the features are extracted using the convolutional layer.Training and testing datasets have been created from the dataset. The model is trained using the training dataset, which makes about 80% of the overall dataset. The trained model is tested using the testing dataset, which makes up 20% of the overall dataset. The training dataset is given to the model after splitting. Then after successfully training the model we pass the testing dataset to the model. Then the model detects the explicit and non-explicit images which is given by the user.

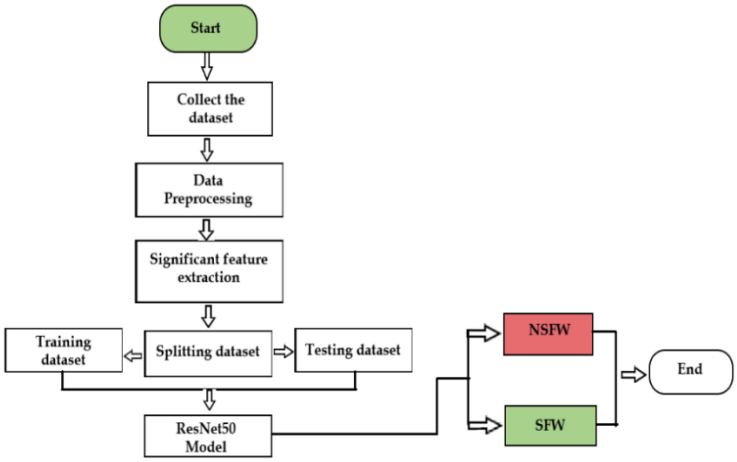

Fig 1.1 System Architecture

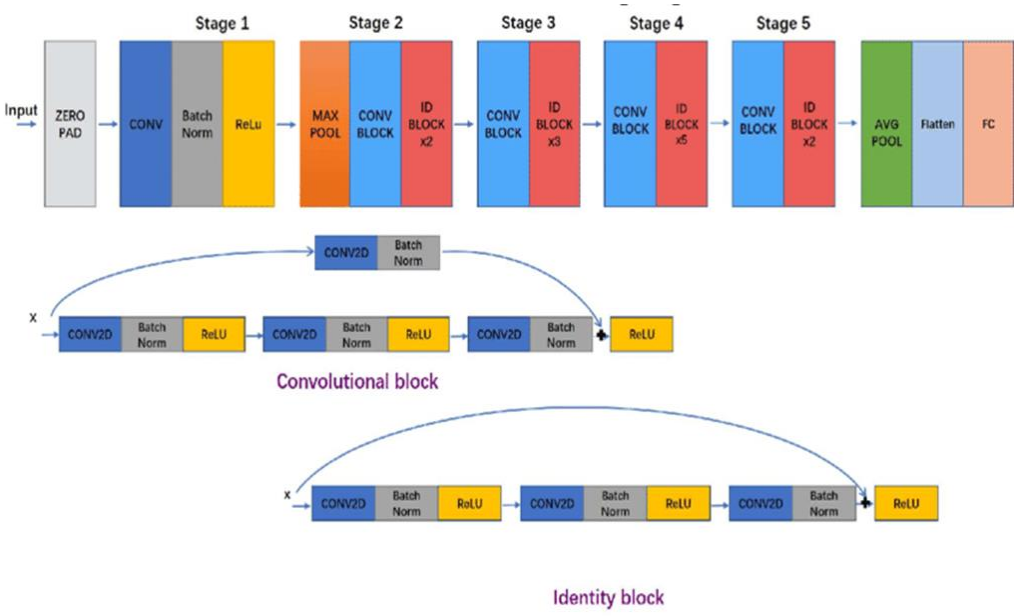

Fig 1.2 ResNet50 Architecture

The above Figure 1.2 the output of one layer is connected to the input of another layer via a skip connection in the ResNet-50 design. I often "skip" two to three layers at a time. Residual networks facilitate the propagation of data from earlier layers to later layers, effectively solving the vanishing gradient problem. We built an implementation of the second model around the popular ResNet50. This implementation used a combination of identity blocks and convolution blocks. A model's identity block consists of window size 1x1, step 1, convolutional layer of ReLu activation, window size FxF, step 1, convolutional layer of ReLu activation, window size 1x1, step 1, another convolutional layer of ReLu activation function. The mid-level FxF window size is uniquely defined for each block. The input of the first convolutional layer is added element by element before the output of the last convolutional layer feeds each ReLu activation. Also, to speed up the learning rate, the output of each convolutional layer is batch normalized before it is fed to its respective ReLu enable. The convolution block in this model consists of the same set of three layers as the identity block. The only difference is that the input of the first layer is convolved with the same convolution before being added element wise to the output of the last convolution. It is batch normalized as a third layer. Taking into account these blocks, the complete network is made up of convolutional layers and blocks. After two identity blocks, one convolutional block, three identity blocks, one convolutional block, five identity blocks, and one convolutional block, a fully connected layer is the last stage. Fully connected layers were tested with one exit for NSFW/SFW classification or four exits for porn/gore/weapons/SFW classification.

5. Conclusion

In this proposed system we have used the ResNet-50 algorithm for image detection and classify it as safe and not safe for work. Our model is based on ResNet-50 algorithm which can give better accuracy. The ResNet-50 architecture provides best result than other architecture, this ResNet-50 requires only some minor modifications. Essentially our model cannot work with the large size images. We would also like to work with the large image size which includes large number of features.

References

[1] Lucas Ege, Isaac Westlund “Building A Not Safe forWork Classifier”(Winter 2018).

[2] Ali Qamar Bhatti, Muhammad Umer, Syed Hasan Adil, Mansoor Ebrahim, Daniyal Nawaz, Faizan Ahmed “Explicit Content Detection System: An Approach Towards A Safe And Ethical Environment” (July 2018).

[3] Daniel Bicho, Artur Ferreira, Nuno Datia “A Deep Learning Approach to Identify NOt Suitable For Work Images” (2020).

[4] Junren Chen, Gang Liang, Wenbo He, Chun Xu, Jin Yang, Ruihang Liu “A pornographic images Recognisation Model Based on Deep One - Class Classification with Visual Attention Mechanism” (Jul 2020).

[5] Agrawal, S., Deshmukh, S., Rawade, R., Desai, M., & Deshmukh, P. (2016). Smart application for food donation using cloud computing. International Journal, 4(4), 683-684

[6] Dmitry Zhelonkin, Nikolay Karpov “Training Effective Model for Real-Time Detection of NSFW Photos” (Feb 2020).Rasoul Banaeeyan, Hezerul Abdul Karim, Haris Lye, Mohamad Faizal Ahmad Fauzi, Sarina Mansor, John See “Automated Nudity Recognition using Very Deep Residual LearningNetwork” (Oct 2019)

[7] S. Agrawal, S. Suryawanshi, V. Arsude, N. Maid and M. Kawarkhe, "Factors Involved in Artificial Intelligence-based Automated HTML Code Generation Tool," 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), Aurangabad, India, 2020, pp. 238-241,doi: 10.1109/ICSIDEMPC49020.2020.9299609.

[8] Agrawal, S. A., Suryawanshi, S., Arsude, V., & Maid, N. (2020). Artificial intelligence based automated HTML code generation tool using design mockups. Journal of Interdisciplinary Cycle Research, 12(III).

[9] Dmirty Zhelonkin, Nikolay Karpov “Training Effective Model for Real-Time Detection of NSFW Photos and Drawings”(2020).

[10] Susanna Paasonen,Kylie Jarrett,B.Light “NSFW: Sex,Humor, and Risk in Social Media” (Oct 2019).

[11] Kawarkhe, M. B., & Agrawal, S. (2019). Smart Water monitoring system using IoT at home. IOSR J. Comput. Eng, 21(1).

[12] Jay Mahadedokar,Gerry Pesavento “Open SOurcing a Deep Learning Solution for detecting NSFW images”.

[13] Kumar, P. (2021). A Proposed Methodology to Mitigate the Ransomware Attack. In Recent Trends in Intensive Computing (pp. 16-21). IOS Press.

[14] Salunke, M., Kabra, R., & Kumar, A. (2015). Layered architecture for DoS attack detection system by combined approach of Naive Bayes and Improved K-means Clustering Algorithm.

International Research Journal of Engineering and Technology, 2(3), 372-377.

[15] Salunke, M. D., & Kabra, R. (2014). Denial-of-service attack detection. International Journal of Innovative Research in Advanced Engineering, 1(11),16-20.

[16] Agrawal, S. A., & Chavan, S. B. (2014). EMS: An android application for emergency patients. Internat J Computer Sci Information Tech, 5(4), 5536-8.

[17] Jadhav, S. B. ., & Kodavade, D. V. . (2023). Enhancing Flight Delay Prediction through Feature Engineering in Machine Learning Classifiers: A Real Time Data Streams Case Study. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2s), 212–218. https://doi.org/10.17762/ijritcc.v11i2s.6064

[18] Wilson, T., Johnson, M., Gonzalez, L., Rodriguez, L., & Silva, A. Machine Learning Techniques for Engineering Workforce Management. Kuwait Journal of Machine Learning, 1(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/120

[19] Vadivu, N. S., Gupta, G., Naveed, Q. N., Rasheed, T., Singh, S. K., & Dhabliya, D. (2022). Correlation-based mutual information model for analysis of lung cancer CT image. BioMed Research International, 2022, 6451770. doi:10.1155/2022/6451770