International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I11P231
Volume 69 Issue 11, 248-255, November, 2021

Original Article

Porn Detection in a Video Streaming Using Hybrid Network of CNN and LSTM

Ilham Bintang, Gede Putra Kusuma

Computer Science Department, Binus Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480

ilham.bintang@binus.ac.id, inegara@binus.edu

Abstract *— Porn detection in video streaming needs an efficient way to recognize because it consists of many picture frames that are stitched together to form a movement. Using real-time frame per frame detection is expensive. On the other hand, using fewer frames will lead to the loss of content. Choosing the right n frame to recognize is good, but it will calculate everything from scratch. A great trick to handle that is to use the information from the previous frame to calculate the feature of the next frame in the sequence. One of the most used approaches to process sequential data is long short-term memory (LSTM). In this research, CNN is combined to reduce the feature complexity and feature extraction and LSTM to store previous frame information to calculate the next frame. For the CNN layer, there are 3 types of models: ResNet50, VGG16, Simple CNN. The ResNet50 model can achieve the best accuracy of 98%. However, the best average inference time is achieved by Simple CNN at 90 ms for a 5-second video.*

Keyword — *Hybrid Network, CNN model, LSTM model, Porn Recognition, Video Streaming*

I. INTRODUCTION

The development of algorithms and technology of computer vision is proliferating. Computer vision application and research have expanded to various scientific fields such as agriculture, education, military, and social. Some applications driven by computer vision are commonly used, and their impact can feel their impact, such as face filter applications on social media, attendance with face recognition, vehicle plate detection, etc. This is possible because of the computational capabilities supported by the GPU (graphical processing unit), which allows for repeated high-dimensional matrix computations [1].

The development of computer vision gradually to be used to solve social problems such as video surveillance, online traffic tickets, biometric security, etc. There is research about real-time violence monitoring by using Deep Learning [2]. On the other hand, many social problems could be solved by using computer vision capabilities—various online platforms providing video streaming, which opens opportunities for irresponsible parties to distribute pornographic content. Manual content quality control in billion hours of video watched every day at YouTube is impossible.

Illegal access to pornographic content among teenagers is not only caused by curiosity alone. However, this can be a form of self-expression that is wrong. Teenagers will think of it as a form of trend and keep it up to date. Worse, this phenomenon will be imitated by children as well. If efforts are not made to limit the content to adolescents and children, in the long run, it will have an impact on the decline in moral quality, character, and character. Concretely, there will be an increase in criminal cases related to immoral acts [3].

Pornographic content can be spread through porn sites, social media attachments or e-mails, and even broadcast live via video streaming platforms. This certainly can affect the image of the video streaming platform. With easy access to pornographic content, service providers often block accounts indicated to reveal pornography or porn action. The problem is, the platform cannot automatically filter the content that is running, so it requires the help of artificial intelligence technology.

Video streaming consists of many frames of a picture that are combined to form a movement. A video file consists of an average of 30 FPS (frames per second) for a typical use case. Using real-time frame-to-frame recognition for video streaming takes time and is an exhaustive computational process. On the other hand, using fewer frames should make the recognition faster but cause a loss of important information, which can reduce the recognition effectiveness—the more frame to recognize, cause the considerable time and computational process [4].

In the context of detecting porn scenes in video streaming, a porn scene is represented as a sequence of frames that contain porn content that probably has a different background, skin colour, pose, and dress colour. Using traditional feature engineering should lead to overfitting because the actual data have a vast variety. Traditional edge detection, colour recognition, background detection will work if there is a lot of data representation for each use case. As mentioned in previous research, CNN and transfer learning are battle-tested workhorses. CNN is a method to automatically look for the unique feature by combining some convolutional layers [5]. Transfer learning will use less data

because it's already trained using many examples and needs to finetune that with the dataset. Also, CNN will reduce the redundancy and complexity when extracting a feature from each frame because there is already a pooling layer and flatten it to the dense layer.

As far as this paper was written, the most recent research of porn recognition was using a combination of CNN and random sampling to recognize frame by frame. The effectiveness of this approach is highly dependent on the random sampling method [5]. Other than that, the results of this study lead to high false positives because when there are few porn frames on video, it will be classified as porn.

Video analysis and video detector technology are proliferating for various cases such as sports video recognition, vehicle counter, and motion detector. For some cases, such as action recognition, a public dataset such as HMDB, UCF, etch is used to find the best approach for the case. However, in the case of the porn video dataset, there is no public dataset that can be used as a benchmark, so that there is a gap from the community need and adequate research to solve it.

Much research related to porn recognition has been conducted because of the impact on the community, especially for Asian culture. So far, the research only covers image recognition, and only one has examined videos, which uses random sampling. That technique will randomly select the frame of video and calculate the average. This method will not work well for video streaming because streaming does not have an exact length. The system should check while the video still streaming [6].

The quick solution to solve real-time video recognition is to check every n frame, adjust n until it has an optimal accuracy; this will work well but need to recompute every n frame. The approach will solve the issue by combining feature extraction from CNN and using the LSTM benefit, which can use previous data as input for the following data. This will use previous frame information as consideration, and the next frame did not compute from scratch.

II. RELATED WORKS

Many types of research related to the introduction of porn content have been carried out, considering the widespread and unsettling impact on the community. So far, the research conducted only covers the introduction of pornographic images, and no one has examined porn videos. Pornographic videos have become a problem for all levels of society because of their very rapid circulation [7].

One of the studies on the introduction of pornographic content is based on the skin appearance probability using the Eigen-porn feature extractor and using the Principal Component Analysis (PCA) method in the YCbCr colour space [8]. This study classifies images containing human genitalia and looks for the Region of Interest (ROI) in the image. The study succeeded in recognizing pornographic content with high accuracy, reaching 90.13%. However, the inference time required to recognize a pornographic image is around 0.12 seconds.

A study with a similar algorithm to detect pornographic images using the YCbCr colour space as skin detection if the image with significant percentage of skin appearance is pornographic images [9]. The detection effectively recognises the skin but finds some errors due to the lighting condition images when taken, other factors that can be the poor interpretation of the system. This study produces the highest accuracy, namely 64.3%, at a threshold of 50.

Also, there is research on pornographic images based on image information content. This research uses colour information and image characteristics (image signature) obtained by the colour histogram technique and wavelet transformation [10]. The test results of this study show that the success rate of the introduction of pornographic images using this method is 67.02% (can recognize as many as 84 images as pornographic images out of 125 tested pornographic images) and detect 36 non-pornographic images into porn. from 375 non-pornographic images (9.06%).

Research on pornographic content has also been conducted by Napa et al. By approaching the percentage of skin and face colour. This study detects child pornographic images by identifying natural person skin colour in images, extracting its features to give unique information of explicit content, and performing age classifications based on facial images. This technique relies heavily on using a skin tone checker and sets of facial features to improve the identification of a child's face. Tests on a dataset containing explicit images captured at different light levels and reflecting the diversity of human skin tones showed an accuracy of about 90% [11].

Other studies about the introduction of porn videos using the CNN method have been conducted before. However, they still carry out a random sampling process in identifying porn videos. So that there will be the possibility of frames that do not enter the recognition process, this study has an accuracy of 100% in the context of the video and 80% for the entire image sample taken [5].

A Combination of CNN and LSTM was used to develop fully integrated violence detection and control by mobile apps. This research has an accuracy of 96.55% for hockey datasets and 98.32% for movie datasets [2].

Previously there was also research on introducing sports activities in videos that combined the CNN and LSTM methods. This study succeeded in obtaining an average accuracy of 92.65% for several actions in the video, such as Basketball, Biking, Diving, Golf-Swing, Soccer, Horse-Riding, Tennis, Volleyball, Swing, Walking, and Jumping. Starting from the success of this method in recognizing actions in the video, it should be able to have a satisfactory performance also in the context of pornography [12].

The most recent approach for video action recognition was using combination CNN + ResNet50 + ConvLSTM[13] and BERT video to encode visual features including spatial,

audio, motion and temporal context in the live video [14]. CNN + ResNet50 + ConvLSTMhas achieved 97-100% of accuracy by using Hockey Fight Dataset, Movies Dataset, and Violent-Flow's dataset [13]. Meanwhile, the BERT feature has achieved69.24% accuracy.

In general, research on pornographic videos has been conducted by several researchers with good enough results to identify pornographic videos. However, the approach taken is still very manual to find out the features of pornographic videos. Additionally, no research has specifically focused on real cases such as video streaming on online platforms such as Facebook or YouTube. Therefore, in the video streaming context, object detection should be performed before classification.

III. THEORY AND METHOD

A. Convolutional Neural Network (CNN)

Convolutional Neural Network is the subsequent development of Multilayer Perceptron (MLP) and is designed to learn from data patterns by convoluting several chunks of images into the various kernel. The convolution operation is the matrix product between the original image and kernel. Kernel is a filter matrix in numbers used to modify the original image[15].

This calculation continues until the kernel centre point has finished accessing all the original image indexes. The convolution operation on a 2-dimensional image can be mathematically calculated using Equation (3-1).[16]

$$S_{(i,j)} = (K * I)_{(i,j)} = \sum\sum I_{(i-m,\ j-n)} K_{(m,\ n)} \quad (3.1)$$

The CNN method has proven to be successful in outperforming other traditional machine learning methods such as SVM in the case of object classification in images. The way CNN works is similar to MLP, but in CNN, each neuron is represented in two dimensions, unlike MLP, where each neuron is only one dimension.

Convolution neural network is one type of neural network that can be used for image classification. Convolution neural network works with several stages where the input and output of each stage consist of several feature maps. Each stage consists of three layers: the convolution layer, the activation function layer, and the pooling layer.

B. ResNet50

ResNet or Residual Network is a residual network that has a deep level. The most profound network of ResNet is about 152 layers. This network is about 8 times deeper than the VGG network, but the complexity is still lower than the VGG network. In 2015, this network won the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) and COCO competitions for the category of image classification, detection and segmentation on COCO and ImageNet datasets[17].

An architecture shows that increasing depth in neural networks generally does not always result in better accuracy; increasing depth can cause a decrease in accuracy when training neural networks. This happens because there is degradation. After all, the network cannot be optimized. The solution is to have the identity mapping and other network layers copied from the shallow model. Thus was born the residual network or Residual Network. This network leaves the layer according to its mapping.

In simple terms, this network is a feed-forward network with a connection/shortcut connected to the following network. This connection has the mapping, and its output is added to the output of another layer. This connection adds no parameters or computational complexity. In addition, this connection can be trained by SGD (Stochastic Gradient Descent) and is easy to implement[17].

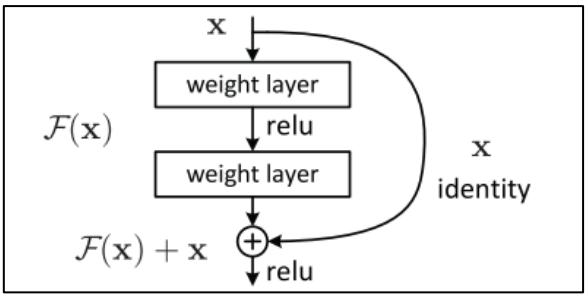


FIGURE 1. ResNet building block

In general, in Figure 1, there is a connection from one layer to the next that can be used directly when the input and output have the exact dimensions. When the dimensions increase, there are two options that can be done, namely mapping with additional dimensions that are 0 or doing a dimension matching process with 1x1 convolution.

C. VGG16

VGG-net created from the Oxford Visual Geometry Group has a layer depth of 16 and 19 called VGG-16 and VGG-19, respectively. This network has 3x3 convolution layers, in addition to the max-pooling layer, which is used to reduce the volume size, and the last layer is fully-connected with 4096 neurons, at the end of which there is a softmax layer. Preprocessing is done on the input by subtracting the average RGB value of each pixel. The pooling process is carried out by max-pooling and is accompanied by several convolution layers [18].

VGG-16 is a model that consists of 16 layers and is fully connected, which is usually used to recognize and classify images. The VGG-16 architecture has 13 convolution layers that use 3x3 filters with a max-pooling layer for downsampling. In addition, there is also a fully connected layer with 4096 layers units followed by a dense layer with 1000 units.

D. LSTM

Long Short-Term Memory (LSTM) is an evolution of the Recurrent Neural Network architecture, which Hochreiter first introduced. Until this research was carried out, many

researchers continued to develop the LSTM architecture in various fields such as video recognition, speech recognition and forecasting. To understand the basic concepts of LSTM, it is necessary to first understand the RNN architecture.

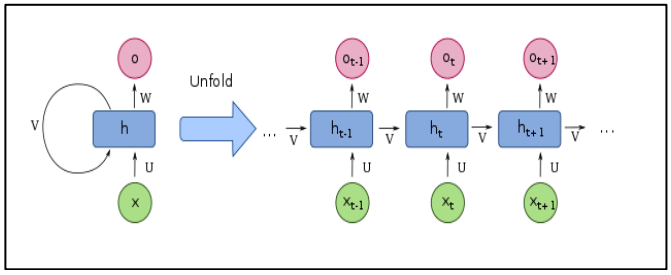


FIGURE 2. Recurrent Neural Network building block

Figure 2 explains that RNN is an information chain that makes the output of the t-the node become the input for the t+1 node. Primarily RNN is used for sequential data, such as air quality monitoring,sing time series data of air quality in an area [19]. RNN has a disadvantage, namely at input X0, X1 has an extensive range of information with Xt, Xt-1 so that when ht+1 requires information relevant to X0, X1 RNN cannot learn to relate information because the old memory stored will be increasingly useless as time goes by because it is overwritten or replaced by new memory. In contrast to RNN, LSTM does not have this drawback because LSTM can manage memory at each input by using memory cells and gate units.

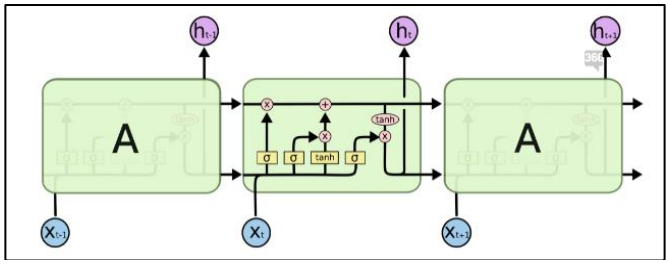


FIGURE 3. LSTM building block

Figure 3 explains how the workflow of memory cells on each LSTM neuron works. There are four activation function processes at each input to the neuron, from now on referred to as gate units. The gate units are forgotten gates, input gates, cell gates, and output gates. In forget gates, the information on each input data will be processed and which data will be stored or discarded in memory cells. The activation function used in these forget gates is the sigmoid activation function. Where the output is between 0 and 1, if the output is 1, then all data will be stored, and vice versa; if the output is 0, then all data will be discarded.

IV. PROPOSED METHOD

In general, the power of combination from the varied architecture of CNN and LSTM [7]. The advantage of CNN is to reduce image redundancy and complexity into sort of feature. Then LSTM is used to pass previous features to be calculated for the next frame until the last frame and be the input for the dense layer. By using a different type of CNN architecture, the effect from different feature extractors should be shown.

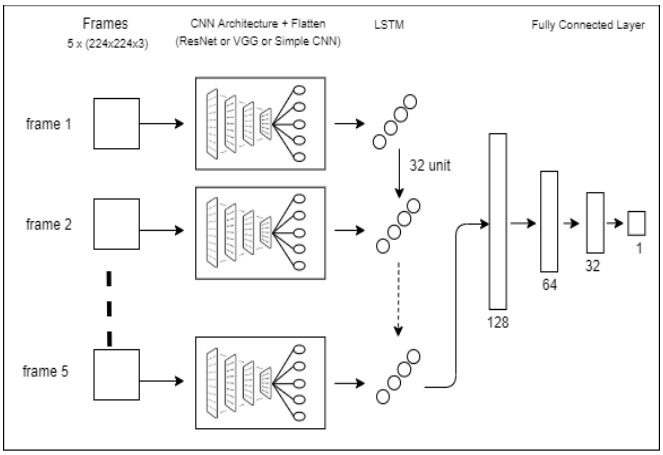


FIGURE4. Architecture hybrid network CNN and LSTM

Figure 4 contains network architecture based on the architecture presented in the paper [12] with some modifications: 1) did not use the image subtraction layer; 2) Flatten last CNN output instead of passing the 2D value of CNN to LSTM; 3) make the dense layer simpler. Assuming porn video recognition needs the whole frame from a porn video, there is no need to use the "subtraction motion" extracted by subtracting frame n + 1 with frame n on the original paper. The combination of CNN's and LSTM will be able to extract sequence features and pass previous vital features to the next frame and dense layer to estimate the class output. As mentioned in Figure 4, there are 3 types of CNN.

CNN is a set of neural networks that are used to extract the representation and classification of images. In the case of real-time recognition in video streaming, every frame feature represents the CNN features, the sequential information between them followed by finding using LSTM. As mentioned in section 2, a video is a set of images/frames that move at ~30 FPS. Each frame on interval FPS might have very similar/redundant frames; processing the whole frames is an expensive computational process. Considering the computational power complexity needed to run the training process, five frames as a video chunk when processing a video for porn recognition. A porn action video is moving from frame to frame which a slight differential in frame and orientation can be observed. As the CNN feature will find the unique patterns from the image/frame, it will notice all the tiny changes. These tiny changes from the sequential video are learned through LSTM.

TABLE1. Simple CNN architecture

#	Layers	Output shape	No.of Param
0	Input layer	5 x 224 x 224 x 3	0
1	Convolutional 2D Stride: 1x1 Padding: 0 Kernel: 3x3 Filters: 128	5 x 222 x 222 x 128	3.584
2	Convolutional 2D Stride: 1x1 Padding: 0 Kernel: 3x3 Filters: 64	5 x 220 x 220 x 64	73.792
3	MaxPooling 2D Kernel: 2x2	5 x 110 x 110 x 64	0
4	Convolutional 2D Stride: 1x1 Padding: 0 Kernel: 3x3 Filters: 64	5 x 108 x 108 x 64	36.924
5	Convolutional 2D Stride: 1x1 Padding: 0 Kernel: 3x3 Filters: 32	5 x 106 x 106 x 32	18.464
6	MaxPooling 2D Kernel: 2x2	5 x 53 x 53 x 32	0
7	Flatten	5 x 89.888	0

Table 1 shows that the input for simple CNN is 5 frame images with 3 colour dimensions with size 224x224. Because of that, the input size is written 5x224x224x3. The Convolutional 2D will treat the input as time distributed that is able to process 5 images parallel. The output from a superficial CNN layer with 5 x 89.888 dimensions will be passed into LSTM with 32 cell units. The differences between the 3 models of CNN on the feature engineering side. So, all models will use the same LSTM architecture.

Analyzing the tiny patterns change in temporal sequential and spatial sequential [12]. Real-time video streaming is sequential data in every change from frame-to-frame are helps the neural networks understand the video context. LSTM can understand such changes in sequences but forget the earlier frame of the sequence data. LSTM is introduced for forgetting the earlier input and is well known as vanishing gradient problem and used to solve the particular type of RNN.

The ResNet50 architecture is referenced from the original paper [17] and VGG16 also from the original paper [18]. The CNN is following some customization to fit the dataset shape. In this case, both ResNet50 and VGG16 are used without making the layer trainable, and the architecture is kept as is. Also, the pretrain weight such as ImageNet is ignored. As mentioned above, all CNN model that proposed is connected to the same LSTM model.

TABLE2. LSTM Architecture

#	Layers	Output Shape	No. of Param
0	Input LSTM Size: 32	32	11.509.888
1	Fully connected Size: 128	128	4.224
2	Fully connected Size: 64	64	8.256
3	Fully connected Size: 256	256	16.640
4	Fully connected Size: 32	32	8.224
5	Dropout Value: 0.1	32	0
6	Output fully connected Size: 1 Activation: Sigmoid	1	33

Table 2 shows input from the LSTM layer are flattened features from CNN models. The LSTM layer is used to pass the current feature to the next frame. The fully connected dense layer is used to reduce the feature complexity on the last frame and used to decide the class, which is porn or non-porn. Output from fully connected dense is 1 layer because the activation is using sigmoid. The calculation will use a single float value to classify the porn or non-porn, and the confidence is scaled from the sigmoid output value.

V. EXPERIMENT

A. Dataset

The dataset used from the KIA dataset and the dataset of the Unram PSTI AI research team is named "PornDbSetTiUnram". Regarding permission to use the data, the author has received written permission from the related party. If necessary, crawling video data from the internet will be carried out as additional test data to ensure that the system built can recognize pornographic content.

In this study, there were several scenarios carried out, such as repetition of the test and the amount of data used was 100 videos for each class with a duration of about 5 minutes, then the video will be processed as an image and extracted frame per frame (only get around 1 frame per second).

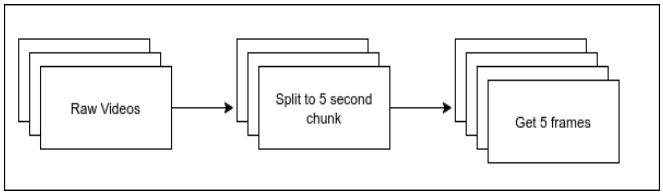


FIGURE 5. Dataset extraction

As shown in Figure 5, the video dataset will separate into images during training progress. Because using 1 frame per second, so the calculation was quite simple. If using 5 seconds of video, then only extract 5 images per video. There are 1000 chunks porn video and 1000 chunks of non-porn videos distributed, as stated in Table 3.

TABLE3. Dataset distribution

Class	Train	Validation	Test
Porn	300	100	100
Non-porn	300	100	100
Total	600	200	200

B. Experimental Design

The evaluation process carried out in this study uses the sampling method from the test data. One by one, the data in the test data becomes a video query. Each video query will be tested into the recognition system. The observation results from all queries will be calculated into the accuracy, recall, precision functions.

The metric calculation checks the quality of the machine learning model. In this case, the Recall, Precision, Accuracy, and F1 score are calculated. Also, measuring the inference time for every CNN model that proposed. This process is carried out as a measure of evaluation in a system. Measuring the level of accuracy can use various ways, one of which is using the Confusion Matrix.

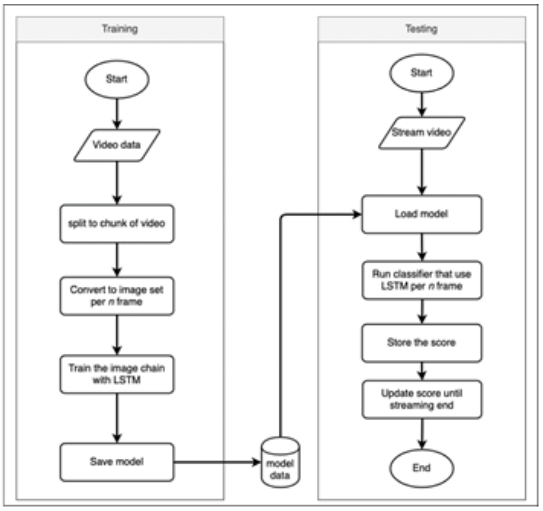


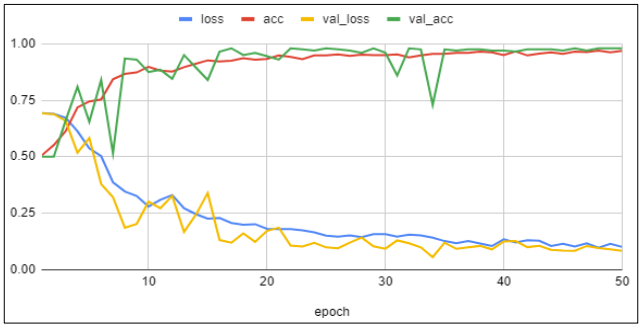
FIGURE 6. Experimental design

From a higher point of view, the whole recognition system is presented in Figure 6, which has two phases: the training phase and the inference phase. The training phase is a process to develop the model. This includes the preparing dataset splitting the video into a trainable image captured every second and the training process itself. This will be used for LSTM to store as a set of images. During the training process, a dataset that has been prepared before will feed to CNN architecture, and the output will be stored to LSTM. The LSTM will pass the stored output as input for the next frame, so on and forth. For the post-processing, the trained model will be stored as a binary file to be loaded for the inference phase.

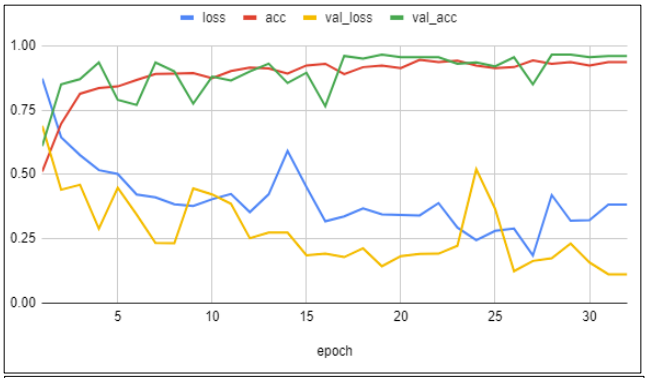
The inference phase determines a query of video streaming as porn or not porn in real-time. The inference process starts with selecting per n frames. Each frame will feed to CNN and LSTM; the result will show and be calculated every n frame. The output is in a streaming processing, so if this system was implemented, the client can consume/subscribe to the streaming output.

C. Experimental Results

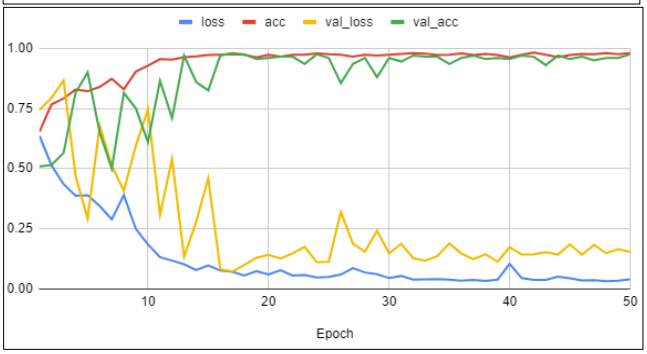
The best model from each experiment comes from hyperparameter tuning and the best experiment following this graph.



(A) ResNet50 Graph



(B) VGG16 Graph



(C) CNN Simple CNN

FIGURE 7. Training graph for each model (A) ResNet50 (B) VGG16 (C) Simple-CNN

Figure 7 shows the best-tuned hyperparameter for each model. ResNet50 training graph is relatively stable because the training accuracy and validation accuracy are similar and did not indicate any overfitting. And VGG16 is quite fluctuating because there are some different values from training and validation accuracy. The last model also immensely fluctuated at the first 20 epochs and became more stable after that. The detailed hyperparameter tuning results are shown in Table 4.

TABLE4. Best hyperparameter

Parameter	ResNet50	VGG16	Simple
No. of epoch	50	46	22
Loss	0.10	0.07	0.07
Accuracy	0.7	0.97	0.98
Val. loss	0.08	0.10	0.16
Val. accuracy	0.98	0.97	0.98
Max Epochs	50	100	100
Learning rate	0.01	0.001	0.0001
Optimizer	SGD	Adam	RMSProp

The hyper-tunning process allows us to find the best performing parameters of the network based on the dataset. The chosen architecture is already presented in Section 3. Table 4 presents the hyper-tunning test accuracy for each of the hyper-parameter values. Also, check the accuracy using test data, which are summarized in Table 5.

TABLE 5. Model performance result

Parameter	ResNet50 (%)	VGG16 (%)	Simple CNN (%)
Accuracy	98.0	94.0	95.5
Precision	100.0	97.82	97.89
Sensitivity	96	90.0	93.0
F1	97.95	93.75	95.38
Average Speed (ms)	92	99	72

Table 5 shows that the significant impact of changing the CNN pre-train layer. In this experiment, there are 200 videos to test the model accuracy and use the best model with the highest validation accuracy from each method. The result shows that the accuracy achieves > 90% for all approaches. These results are probably achieved by providing a good dataset and video chunk. Other metrics that are calculated are the average time to predict 200 video chunks. This shows that Resnet50 achieves the best time and accuracy to predict.

VI. CONCLUSION AND FUTURE WORK

The highest average confidence level can be achieved by using Resnet50 (1e-2 learning rate, 50 epochs, and SGD optimizer). The lowest average inference time could achieve by using Simple CNN with 141 ms. The accuracy by using Simple-CNN is 96% and F1 95.25%. The best accuracy and F1 score method is ResNet50 with 98% accuracy and 97.95% F1 score. ResNet50 has a precision of 100% and a sensitivity of 96%. VGG16 is not so good in accuracy and speed, 94% accuracy, and 99ms average inference speed. The method recommended for video streaming recognition is using ResNet50. The next improvement should be handling the video in real-time and using more data to represent the variation of porn videos such as different poses, the background of the pornographic contents in the video, and the other sexual content.

ACKNOWLEDGMENT

The dataset is partially supported by Tim Riset AI PSTI Unram (Universitas Mataram). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] M. Perez et al., Video pornography detection through deep learning techniques and motion information, Neurocomputing, 230 (2017) 279–293, doi: 10.1016/j.neucom.2016.12.017.

[2] S. Sharma, B. Sudharsan, S. Naraharisetti, V. Trehan, and K. Jayavel, A fully integrated violence detection system using CNN and LSTM, International Journal of Electrical and Computer Engineering, 11(4) 3374–3380, Aug. 2021, doi: 10.11591/ijece.v11i4.pp3374-3380.

[3] G. Dines, Growing Up with Porn: The Developmental and Societal Impact of Pornography on Children, Dignity: A Journal on Sexual Exploitation and Violence, 2(3) (2017), doi: 10.23860/dignity.2017.02.03.03.

[4] Q. Lan, Z. Wang, M. Wen, C. Zhang, and Y. Wang, High Performance Implementation of 3D Convolutional Neural Networks on a GPU, Computational Intelligence and Neuroscience, 2017(2017), doi: 10.1155/2017/8348671.

[5] I. W. A. Arimbawa, I. G. P. S. Wijaya, and I. Bintang, Comparison of simple and stratified random sampling on porn videos recognition using CNN, 2019 International Conference on Computer Engineering, Network, and Intelligent Multimedia, CENIM 2019 - Proceeding, 2019 (2019), doi: 10.1109/CENIM48368.2019.8973305.

[6] L. Wang, J. Zhang, Q. Tian, C. Li, and L. Zhuo, Porn Streamer Recognition in Live Video Streaming via Attention-Gated Multimodal Deep Features, IEEE Transactions on Circuits and Systems for Video Technology, 30(12) (2020) 4876–4886, doi: 10.1109/TCSVT.2019.2958871.

[7] M. Zufar and B. Setiyono, Convolutional Neural Networks Untuk Pengenalan Wajah Secara Real-Time, Jurnal Sains dan Seni ITS, 5(2). 128862, 2016, doi: 10.12962/j23373520.v5i2.18854.

[8] I. G. P. S. Wijaya, I. B. K. Widiartha, K. Uchimura, M. S. Iqbal, and A. Y. Husodo, Fast pornographic image recognition using compact holistic features and multi-layer neural network, International Journal of Advances in Intelligent Informatics, 5(2)(2019) 89–100, doi: 10.26555/ijain.v5i2.268.

[9] J. A. M. Basilio, G. A. Torres, G. S. Pérez, L. K. T. Medina, and H. M. P. Meana, Explicit image detection using YCbCr space color model as skin detection, Applications of Mathematics and Computer Engineering - American Conference on Applied Mathematics, AMERICAN-MATH'11, 5th WSEAS International Conference on Computer Engineering and Applications, CEA'11, (2011) 123–128.

[10] I. G. P. S. Wijaya, I. B. K. Widiartha, and S. E. Anjarwani, Pornographic Image Recognition Based on Skin Probability and Eigenporn of Skin ROIs Images, TELKOMNIKA (Telecommunication Comput. Electron. Control), 13(3) 985, 2015.

[11] N. Sae-Bae, X. Sun, H. T. Sencar, and N. D. Memon, Towards automatic detection of child pornography, 2014 IEEE International Conference on Image Processing, ICIP 2014, no. January, (2014) 5332–5336, doi: 10.1109/ICIP.2014.7026079.

[12] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features, IEEE Access, 6 (2018) 1155–1166, 2017, doi: 10.1109/ACCESS.2017.2778011.

[13] S. Sudhakaran and O. Lanz, Learning to detect violent videos using convolutional long short-term memory, IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), no. doi: 10.1109/AVSS.2017.8078468, (2017) 1–6.

[14] L. Wang, J. Zhang, M. Wang, J. Tian, and L. Zhuo, Multilevel fusion of multimodal deep features for porn streamer recognition in live video, Pattern Recognition Letters, 140(2020) 150–157, doi: 10.1016/J.PATREC.2020.09.027.

[15] P. Kim, Convolutional Neural Network, in MATLAB Deep Learning, Berkeley, CA: Apress, 2017. doi: 10.1007/978-1-4842-2845-6_6.

[16] Boki Latupono, Implementasi Deep Learning menggunakan Convolution Neural Network untuk Klasifikasi Gambar," UNIVERSITAS ISLAM INDONESIA, 2018.

[17] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem, (2016) 770–778, doi: 10.1109/CVPR.2016.90.

[18] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, (2015) 1–14.

[19] S. K. Borse and D. v Patil, Air Quality Prediction Using Recurrent Neural Network, International Journal on Emerging Trends in Technology (IJETT), 7(1) (2020).