2024 10th International Conference on Communication and Signal Processing (ICCSP) | 979-8-3503-5306-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCSP60870.2024.10544299

DeepFake Face Detection using Machine Learning with LSTM

Thipparthi Vignesh
IT Department
B V Raju Institute of Technology
Narsapur, India
20211a12b2@bvrit.ac.in

Potharlanka Harish Tarun
IT Department
B V Raju Institute of Technology
Narsapur, India
20211a1296@bvrit.ac.in

Ryagalla Parthav
IT Department
B V Raju Institute of Technology
Narsapur, India
20211a12a0@bvrit.ac.in

V Bhargavi
Assistant professor, IT Department
B V Raju Institute of Technology
Narsapur, India
bhargavi.v@bvrit.ac.in

***Abstract*— Fake face images that are increasingly convincing and realistic can be created because to the development of face image manipulation (FIM) technologies like Face to Face and Deepfake, which can damage the legitimacy and trustworthiness of online content. Malicious uses of these technology include blackmailing people, posing as celebrities, and disseminating false information. As a result, creating trustworthy and strong techniques to identify FIM and safeguard the integrity of digital media is essential. Numerous current techniques utilize on models built on convolutional neural networks (CNNs), which are capable of detecting FIM by examining a face's visual characteristics. But because these models are frequently tested and trained on certain datasets or circumstances. Furthermore, they might not be able to record the temporal information that is included in video data and can be used to identify irregularities or strange anomalies in FIM videos. We provide a novel method that uses both geographical and temporal information to detect FIM in order to get over these difficulties. We present a new type of residual network called CRNet, which is dependent on Convolutional Long Short-Term Memory (LSTM) and is capable of processing a series of consecutive pictures taken from a movie. The model can learn temporal information because to its design, which is essential for spotting oddities that occur in between frames of FIM movies. We performed extensive tests with several kinds of FIM videos from the Kaggle dataset.**

***Index Terms*— Deepfake detection, Long-Short Term Memory (LSTM), Kaggle, Residual next convolution neural network (Xception CNN), Image manipulation**

I. INTRODUCTION

In the field of machine learning, deep learning (DL) computers has emerged as the industry standard. Software that threatens privacy, democracy, and national security has been made possible by it, such as deepfakes.

One person's face or voice is substituted for another person in deepfakes, which are digital media manipulations such as pictures or movies. Making it appear as though someone said or did something they never said or did by using deep learning algorithms to create or modify audio and video content. By imitating a person's words and movements, deepfake technology produces convincing and realistic fake audio or video. It raises moral questions about consent, privacy, and disinformation since it is frequently used for amusing, pornographic, or political objectives. A comprehensive approach is needed to address these issues, including regulatory measures, public awareness, and responsible AI usage to reduce harm. Deepfakes target social media platforms, taking advantage of the ease of spreading conspiracies and misinformation due to users' tendency to follow the crowd. The use of advanced deep neural networks and abundant data has made the fake images and videos hard to distinguish, fooling both humans and computer algorithms. This study promotes the research in deep learning and deepfake detection. The main focus is on Long-Short Term Memory (LSTM), [5] Convolution Neural Network (CNN) algorithm for deepfake detection. In recent years, deep learning-based techniques for generating synthetic images have increased significantly. These techniques can produce lifelike images, often undetectable by the average human eye. While these techniques have applications in various domains, such as human face generation and realistic scenery creation, their misuse has caused serious social problems. People with malicious intent have created fake videos of celebrities and the public. Various approaches have been proposed to counter this, aiming to detect and mitigate the impact of deepfake videos This misuse has resulted in social consequences, with a recent study indicating that most (95%) of deepfakes are based on explicit content, falling under the broader category of Deepfakes.

The researchers have launched several deepfake information, data to assist researchers progress detection methods and algorithms. [6] A notable example is the Kaggle dataset where we took 80% data to the train the system and 20% for testing. Nowadays many companies are conducting a deepfake detection challenges, giving a prize pool of lakhs to encourage research in this field. Despite the emergence of high-performing deepfake detection methods with impressive accuracy on specific training datasets their effectiveness decreases when faced with new deepfake methods absent from the training set. This challenge excited us to develop a dynamic deepfake image and video detector. Recognizing the impracticality of creating datasets for every new deepfake method, [6] we use the vast Kaggle dataset and transfer learning to enhance detection across various deepfake methods, including new ones.

II. RELATED WORK

Our work lies in various domains, for instance detecting deepfake face in image part and video.

A) Deepfake Detection

Deep Fake Detection Abnormal eyelid detection has been shown to be effective in identifying unpredictability in exploited videos or images. Moreover, techniques for detecting image splices try to take advantage of gaps left by splicing close to the edges of the altered image areas. While it is possible to detect unpredictability in images and videos which are made by available deepfake methods. Constantly sophisticated techniques are being investigated and developed annually, and available deepfake generation techniques are being utilized to build them. [2] Instead of, model-based methods, such as analysing the attributes of artifacts, JPEG artifacts and lens aberrations, resulting from the choosing of substitute research methods. To identify image alteration, an image is typically used. Since the whole image set is built from

scrape, they haven't shown themselves to be trustworthy in identifying false machine-generated images produced by Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). Deep learning models are excellently detecting accuracy in a monitored setting. In specific, advancement based on Convolutional Neural Networks (CNN) is mostly concentrated on gaining from hierarchical representations from input RGB colour images automatically or on leveraging tampered images and videos detection and using handcrafted methods.

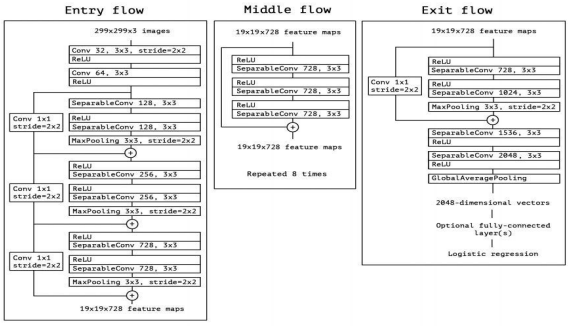

Fig 1: Internal Data flow Architecture

The convolutional short-term memory cell is visually represented. Although some additional parts are included to accommodate the convolution part, the overall structure is comparable to LSTM. In this case, the input is Xt, the cell output is Ct, and the hidden state is Ht. This does not replace portals. Shallow Net, a dynamic and robust learning CNN, to identify GAN-generated images of consisting great precision at 64 x 64 resolution. [4][5] Furthermore, a quicker dual-stream R-CNN network that can record both big-level and small-level visual details is employed. Nevertheless, it is critical to examine the temporal data among successive frames in videos of deepfake. Our Methodology is trying numerous successive frames to leverage this temporary date to perform better and detect deepfakes accurately. Through, this we can enhance the accuracy of the system, which improves its precision.

B) Detection with Consecutive Video Frames

Detection based on sequential video and images can be suggested by a detection procedure that uses [5] CNN (Convolution Neural Network) to catch the secular knowledge consisted in 5 successive frames of a deepfake video. Many researchers have tried detecting but they are fall behind to make result for dynamic images and videos one of them is Guerra took a similar approach, using [5] CNN layers to extract features from up to 50 consecutive images and taking them into Recurrent Neural Network layers to make a deepfake perception model that takes temporal info into account. Both methods observe characteristics from the Convolution Neural Network layers. Correspondingly, we develop CLRNet model employing LSTM convolutional cells, that are able to absorb spatiotemporal knowledge straight from the given input image order. When tested on datasets that included films from a different deepfake generating technique, the majority of these techniques, however, underperformed. To overcome this difficulty, we investigated transfer learning with CLRNet.

C) Generalization via Transfer Learning

Variety of methods given to make deepfake videos that are continuously emerging, and more diversified videos will emerge in coming times. To address this, we use CLRNet.A high-level engineered illustration of LSTM-based CLRNet replica is depicted. The prototype processes a sequence of successive images taken as input and produces a classified output, indicating REAL or FAKE. [1] The layer in the Keras terminology have been utilized. In certain scenarios, employing Multiple Transfer Learning (TL) proves pivotal for finding deepfakes generated through various methods. [2][3] This implies that awareness gathered in one province, like Face manipulation, Face Swapping, Face tampering can enhance generalizability in another province, like

Face-to-Face. Through our project, we present a comparative analysis of our viewpoint with Forensics Transfer, demonstrating superior applicability and consignable.

III. SYSTEM OVERVIEW

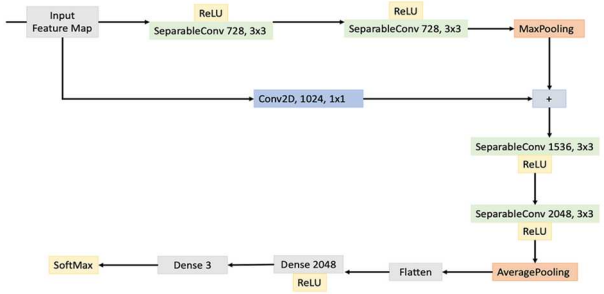

Fig 2: Steps of Data Processing

Database: We use a combined data consisting of number of images and videos originating at divergent data origins such as Snapchat, X, Internet and Kaggle Deep Challenge knowledge for fake content detection. Our reprocessed information contains 50% primary videos followed by 50% exploited videos. The data gained is segregated into 80% coaching and 20% examination sets.

Preprocessing: Preprocessing of taken dataset involves dividing the video content entering frames. Next, face detection happens and the frame with the located face is cropped. To keep the sequential frames consistent, the video in the information is averaged and a up to date progressed dataset is developed, tampering, cropping faces with frames that match this average. Frames that do not contain walls are neglected while progressing. A rough example of Processing a 10-second multimedia at 20 frames/second, or 200 frames in total, requires a large amount of progressing power. Therefore, for examinable reasons, we proposed to use 50% images to instruct the prototype.

Model: The prototype contains of Xception50 in order by an LSTM layer. [4] The information filler fills pre-processed videos with cropped faces, tampered faces and splits them into a instruct batch and a examine batch.

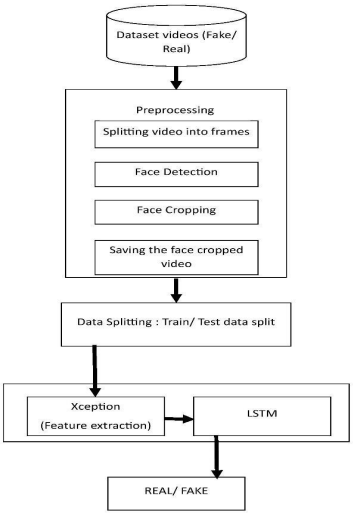

Fig 3: Data Flow Diagram

Xception: Xception categoriser to extract characteristics and precisely detect frame-level features. Furter, we refine the matrix by attaching supplementary essential layers and opt for the appropriate learning rate to successfully intersect with the prototype declination. The characteristic vectors, comprising 2048 dimensions, obtained from the concluding pooling layers, are subsequently employed as sequential input for the Long Short-Term Memory (LSTM) network.

LSTM: In the context of sequence processing, consider a series of input frame feature vectors extracted from a Xception. These vectors, along with a two-node neural network, serve as input, providing the probability that the series corresponds to either a manipulated video or an unaltered video. The primary challenge lies in crafting a model capable of recursively and meaningfully processing these sequences. To address this challenge, we suggest employing an LSTM with 2048 units and a dropout probability of 0.6, effectively achieving our objective. The LSTM is utilized for sequential frame processing, enabling temporal analysis by comparing a frame at time "x" seconds with a frame at "x-y" seconds, where "y" represents any number of frames preceding time "x."

conv_lstm2d_input	input:	[(None, 20, 64, 64, 3)]
InputLayer	output:	[(None, 20, 64, 64, 3)]

conv_lstm2d	input:	(None, 20, 64, 64, 3)
ConvLSTM2D	output:	(None, 20, 62, 62, 4)

max_pooling3d	input:	(None, 20, 62, 62, 4)
MaxPooling3D	output:	(None, 20, 31, 31, 4)

time_distributed(dropout)	input:	(None, 20, 31, 31, 4)
TimeDistributed(Dropout)	output:	(None, 20, 31, 31, 4)

conv_lstm2d_1	input:	(None, 20, 31, 31, 4)
ConvLSTM2D	output:	(None, 20, 29, 29, 8)

max_pooling3d_1	input:	(None, 20, 29, 29, 8)
MaxPooling3D	output:	(None, 20, 15, 15, 8)

time_distributed_1(dropout_1)	input:	(None, 20, 15, 15, 8)
TimeDistributed(Dropout)	output:	(None, 20, 15, 15, 8)

conv_lstm2d_2	input:	(None, 20, 15, 15, 8)
ConvLSTM2D	output:	(None, 20, 13, 13, 14)

max_pooling3d_2	input:	(None, 20, 13, 13, 14)
MaxPooling3D	output:	(None, 20, 7, 7, 14)

time_distributed_2(dropout_2)	input:	(None, 20, 7, 7, 14)
TimeDistributed(Dropout)	output:	(None, 20, 7, 7, 14)

conv_lstm2d_3	input:	(None, 20, 7, 7, 14)
ConvLSTM2D	output:	(None, 20, 5, 5, 16)

max_pooling3d_3	input:	(None, 20, 5, 5, 16)
MaxPooling3D	output:	(None, 20, 3, 3, 16)

flatten	input:	(None, 20, 3, 3, 16)
Flatten	output:	(None, 2880)

dense	input:	(None, 2880)
Dense	output:	(None, 2)

Fig 4: Sequential Processing of Algorithm

IV. EXPERIMENTS AND RESULTS

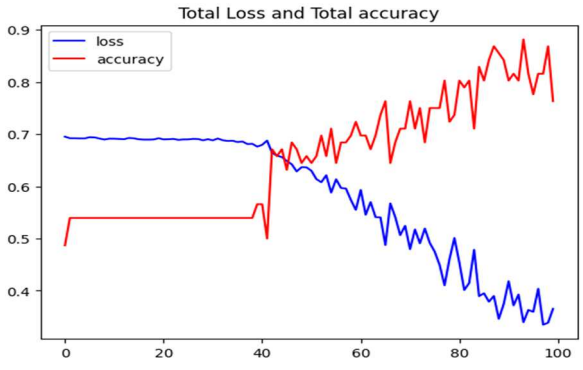

Fig 5: Total Loss and Total Accuracy Graph

Fig 6: Fake Video

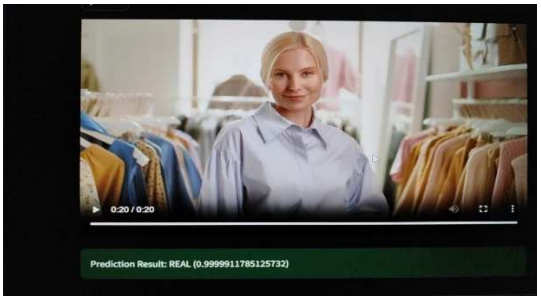

Fig 7: Real Video

Fig 8: Real Image

Fig 9: Training_loss vs Validation_loss

V. CONCLUSION

In conclusion, the presented work focuses on addressing the growing concern of fake face images generated through advanced Face Image Manipulation (FIM) technologies like Deepfake, Face-to-face. These

technologies pose serious threats to online content's legitimacy and trustworthiness, leading to potential malicious uses like identity theft, misinformation, and blackmail.

The paper introduces a novel approach that combines both geographical and temporal knowledge for the detection of FIM, aiming to overcome limitations in existing techniques. The proposed CRNet (Convolutional Long Short-Term Memory) model is designed to process sequential frames from images, videos, leveraging its ability to catch temporal data crucial for identifying anomalies in FIM videos.

The research acknowledges the challenges posed by the dynamic nature of FIM methods, where new techniques constantly emerge. [6] To enhance the model's generalizability across various FIM methods, the paper employs transfer learning, using the Kaggle dataset and leveraging insights gained from one domain to improve generalizability in another.

REFERENCES

[1] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24480–24489, 2023.

[2] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pp. 2316–2324, 2022

[3] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12873–12883, 2021.

[4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG), 42(4):1–10, 2023

[5] Deepfake Detection Challenge. 2019. Deepfake DetectionChallengehttps//www.kaggle.com/c/deepfake-detection-challenge. Accessed: 2020-02-12.

[6] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. 2019. Protecting World Leaders Against Deep Fakes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.

[7] Jawadul H Bappy, Amit K Roy-Chowdhury, Jason Bunk, Lakshmanan Nataraj, and BS Manjunath. 2017. Exploiting spatial structure for localizing manipulated image regions. In Proceedings of the IEEE international conference on computer vision. 4970–4979.

[8] Deepfake Detection Challenge. 2019. Deepfake DetectionChallengehttps://www.kaggle.com/c/deepfake-detection-challenge. Accessed: 2020-02-12.

[9] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184, 2022.

[10] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 22412–22423, October 2023.

[11] Y. Mirsky and W. Lee, “The Creation and Detection of Deepfakes: A Survey,” ACM Computing Surveys, vol. 54, no. 1, pp. 7:1–7:41, 2021.

[15] L. Guarnera, O. Giudice, and S. Battiato, “DeepFake Detection by Analyzing Convolutional Traces,” in CVPRW, 2020, pp. 666–667. [33] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, “DeepfakeDetectionusing Spatiotemporal Convolutional Networks,” arXiv:2006.14749 [cs, eess], 2020.