WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Deepfake Detection and Prevention

Aparna Pandey[1], Ruchi Soni[2], Nitin Kumar Sahu[3], Vanshaj H. Bawane[4], Dennis Marc Zaman[5]

[1]Assistant Professor, Bhilai Institue of Technology, Raipur, Chhattisgarh, India
[2,3,4]Student, Bhilai Institute of Technology, Raipur, Chhattisgarh, India

A B S T R A C T

Our technique can recognise replacement and recreation deep fakes automatically. Our goal is to combat artificial intelligence (AI) by using it against it. Our system extracts frame-level features using a Res-Next Convolution neural network, and then uses these features to train an LSTM-based Recurrent Neural Network (RNN) to classify videos based on whether or not they have been altered, i.e., whether they are deepfake or authentic. To emulate the real time scenarios and make the model perform better on real time data, we evaluate our method on large amount of balanced and mixed dataset prepared by mixing the various available dataset like Face Forensic++ [1], Deepfake detection challenge [2], and Celeb-DF [3]. We also show how our system can achieve competitive result using very simple and robust approach.

Keywords: Res-Next Convolution neural network, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM).

1. Introduction

Deepfakes are terms used to describe multimedia that has been artificially or digitally manufactured or manipulated. Deepfakes are produced through face swapping, acting out or animating facial expressions, using digitally produced sounds, or using fake human faces. Face manipulation, on the other hand, entails changing aspects of the face like age, gender, ethnicity, morphing, attractiveness, skin tone or texture, hair color, style, or length, eyeglass, makeup, moustache, emotion, beard, pose, gaze, mouth open or closed, eye color, injury, and drug use effects, as well as adding imperceptible perturbations.

1.1 Project Idea

It becomes very important to spot the difference between the deepfake and pristine video. We are using AI to fight AI. Deepfakes are created using tools like FaceApp and Face Swap, which using pre-trained neural networks like GAN or Auto encoders for these deepfakes creation. Our method uses a LSTM based artificial neural network to process the sequential temporal analysis of the video frames and pre-trained CNN to extract the frame level features. Convolution neural network extracts the frame-level features, and these features are further used to train the Long Short-Term Memory based artificial Recurrent Neural Network to classify the video as Deepfake or real. The following objectives include:

- To remove biasness in the dataset.
- To create a model which works on diverse dataset collected as single dataset.
- To develop detection and prevention solution

1.2 Tables

Table 1 – Some deepfake tools.

TOOL	DESCRIPTION
Lensa AI	It has ability to create people's bodies in videos.
Deepfake web	It uses deep learning to absorb the various complexities of face data.
Reface	Reface AI uses GAN behind the scenes.
My heritage	It is used to animate old photos.
Deepfake lab	This allows to create deepfake videos. It is primarily built for researchers and students of computer vision.

Table 2 – Datasets for deepfake detection and generation

NAME	DESCRIPTION
Face Forensics++	FaceForensics++ is a forensics dataset consisting of 1000 original video sequences that have been manipulated with four automated face manipulation methods: Deepfakes, Face2Face, FaceSwap and NeuralTextures. The data has been sourced from 977 youtube videos and all videos contain a trackable mostly frontal face without occlusions which enables automated tampering methods to generate realistic forgeries.
Celeb-DF	**Celeb-DF** is a large-scale challenging dataset for deepfake forensics. It includes 590 original videos collected from YouTube with subjects of different ages, ethnic groups and genders, and 5639 corresponding Deepfake videos.
DFDC	The DFDC (Deepfake Detection Challenge) is a dataset for deepfake detection consisting of more than 100,000 videos.
WildDeepfake	**WildDeepfake** is a dataset for real-world deepfakes detection which consists of 7,314 face sequences extracted from 707 deepfake videos that are collected completely from the internet. WildDeepfake is a small dataset that can be used, in addition to existing datasets, to develop more effective detectors against real-world deepfakes.
WaveFake	WaveFake is a dataset for audio deepfake detection. The dataset consists of a large-scale dataset of over 100K generated audio clips.
FakeAVCeleb	FakeAVCeleb is a novel Audio-Video Deepfake dataset that not only contains deepfake videos but respective synthesized cloned audios as well.

1.3 Model Training Details

1. Train Test Split:

The dataset is split into train and test dataset with a ratio of 70% train videos and 30% test videos. The train and test split are a balanced split i.e., 50% of the real and 50% of fake videos in each split.

2. Data Loader:

It is used to load the videos and their labels with a batch size of 4.

3.Adam optimizer:

To enable the adaptive learning rate Adam optimizer with the model parameters is used.

4. Cross Entropy:

To calculate the loss function Cross Entropy approach is used because we are training a classification problem.

5. Softmax Layer:

A Softmax function is a type of squashing function. Squashing functions limit the output of the function into the range 0 to 1. This allows the output to be interpreted directly as a probability. Similarly, softmax functions are multi-class sigmoid, meaning they are used in determining probability of multiple classes at once. Since the outputs of a softmax function can be interpreted as a probability (i.e., They must sum to 1), a softmax layer is typically the final layer used in neural network functions. It is important to note that a softmax layer must have the same number of nodes as the output later. In our case softmax layer has two output nodes i.e., REAL or FAKE, also Softmax layer provide us the confidence(probability) of prediction.

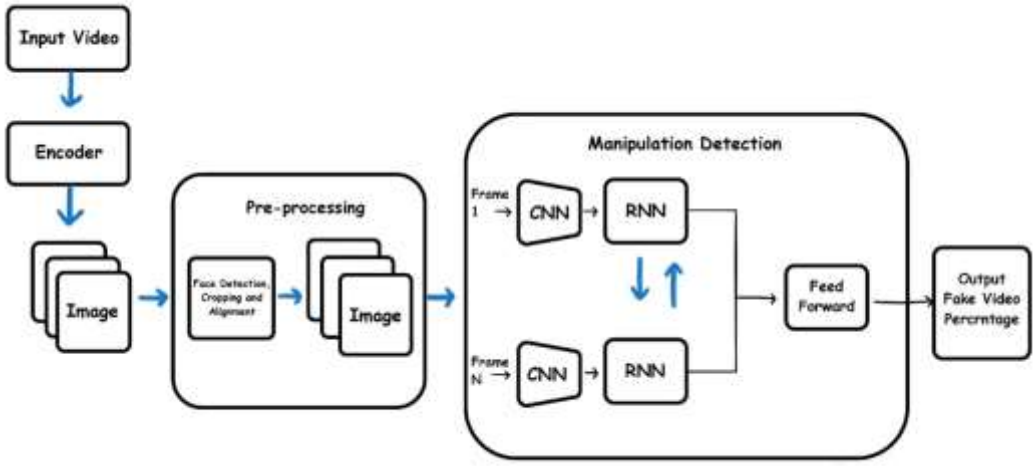
6. Confusion Matrix:

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made. Confusion matrix is used to evaluate our model and calculate the accuracy.

7. Export Model:

After the model is trained, we have exported the model. So that it can be used for prediction on real time data.

2. Methodology



3. Results

One of the primary expected outcomes is the development of robust and accurate deepfake detection models. These models should be capable of identifying deepfake content with a high degree of accuracy.

The project may also yield prevention strategies or methods that can be employed to reduce the creation and dissemination of deepfakes. This might include recommendations for platform security, content verification, or user education.

The development of user-friendly detection tools or software that can be used by individuals, organizations, or platforms to identify deepfake content.

4. Conclusions

Advancements in deepfake detection have proposed a combination of traditional techniques (CNN, DNN and LTSM) and other modules to strengthen these techniques, producing ensembled and multi-attentional architectures. Recent papers written have adopted this strategy with many suggesting multi-attentional architectures in 2021 papers and papers written in 2022 only suggesting multi-attentional architectures. With the evolution of technology and the low-cost barriers to entry, the advancement of deep fakes will progress with a rapid trajectory. As this evolves, future challenges in detection techniques will be required to adapt at the same level or greater to those technological advancements**.** Also, it is hoped that this survey paper will motivate budding scientists, practitioners, researchers, and engineers to consider deepfakes as their domain of study.

References

[1] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, Yang Liu, “Countering Malicious DeepFakes: Survey, Battleground, and Horizon” in International Journal of Computer Vision (2022) 130:1678–1734.

[2] Zahid Akhtar, “Deepfakes Generation and Detection: A Short Survey” in Journal of Imaging.

[3] Laura Stroebel, Mark Llewellyn, Tricia Hartley, Tsui Shan Ip & Mohiuddin Ahmed (2023) “A systematic literature review on the effectiveness of deepfake detection techniques”, in Journal of Cyber Security Technology, 7:2, 83-113, DOI: 10.1080/23742917.2023.2192888.

[4] Curry L, Nembhard I, Bradley E. Qualitative and mixed methods provide unique contributions to outcomes research. Circulation. 2009;119(10):1442–1452.

[5] Abdulreda AS, Obaid AJ. A landscape view of deepfake techniques and detection methods. Int J Nonlinear Anal Appl. 2022;13(1):745–755.

[6] Hazirbas C, Bitton J, Dolhansky B. Towards Measuring Fairness in AI: the Casual Conversations Datase. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2021.

[7] Weerawardana M, Fernando T, Deepfakes detection methods: a literature survey, in 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS). 2021, IEEE Access: Negambo, Sri Lanka. p. 76–81.

[8] Almutairi Z, Elgibreen H. A review of modern audio deepfake detection methods: challenges and future directions. Algorithms. 2022;15(5):155.