Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 05, May-2025

Deepfake Detection System

Research Paper by

Author's Name(S): - Chandraprakash Sahu

Under the supervision of

Mr. Komal Yadav
Assistant Professor
Supervisor Name

Mr. Kameshwar Rao
Assistant Professor
Supervisor Name

Department of Computer Science & Engineering Faculty of Engineering

Institutional Affiliation:- Shri Rawatpura Sarkar University Raipur

1. Abstract

The proliferation of deep learning algorithms has simplified the creation of highly realistic human-synthesized videos, commonly known as deepfakes. These fabricated videos pose significant threats, with potential applications ranging from political manipulation and fake terrorism events to revenge porn and blackmail. This paper introduces a novel deep learning-based method designed to effectively distinguish AI-generated fake videos from authentic ones. The proposed system is capable of automatically detecting both replacement and reenactment deepfakes. It leverages a Res-Next Convolutional Neural Network (CNN) to extract frame-level features, which are then used to train a Long Short- Term Memory (LSTM) based Recurrent Neural Network (RNN). This combined approach classifies whether a video has undergone manipulation, identifying it as either a deepfake or a real video. To ensure robust performance in real-time scenarios, the method was evaluated on a comprehensive dataset, integrating various existing datasets such as Face- Forensic++, Deepfake detection challenge, and Celeb-DF. The system demonstrates competitive results through its simple and robust methodology.

2. INTRODUCTION

Background Information

The rapid advancements in mobile camera technology and the widespread adoption of social media platforms have made the creation and dissemination of digital videos more accessible than ever before. Concurrently, deep learning has enabled the development of technologies previously deemed impossible, including sophisticated generative models capable of synthesizing highly realistic images, speech, music, and video. While these generative models have beneficial applications, such as text-to-speech for accessibility and generating training data for medical imaging, they have also given rise to significant challenges. One such challenge is the emergence of "deepfakes," which are manipulated video and audio clips produced by deep generative models. Since their initial appearance in late 2017, numerous open-source deepfake generation methods and tools have emerged, contributing to a growing volume of synthesized media. Although some deepfakes are created for humorous purposes, many others carry the potential for harm to individuals and society by spreading misinformation and creating distress. The increasing realism and accessibility of deepfake creation tools necessitate robust detection mechanisms.

Research Problem

The ease of creating deepfakes using artificially intelligent tools contrasts sharply with the difficulty of detecting them, posing a major challenge. Historical instances demonstrate the power of deepfakes in generating political tension, fabricating terrorism events, enabling revenge porn, and facilitating blackmail. Therefore, developing effective methods for deepfake detection is crucial to prevent the unchecked proliferation of these manipulated media across social media platforms.

Objectives

This project aims to address the critical need for deepfake detection through the following objectives:

- To expose the distorted truth behind deepfakes.
- To mitigate the spread of abuses and misleading information to the public on the internet.
- To accurately distinguish and classify videos as either deepfake or pristine.
- To provide a user-friendly system that allows users to upload videos and determine their authenticity.

Thesis Statement

This research proposes and validates a deep learning-based system that effectively combats AI-generated deepfakes by leveraging the inherent artifacts left by deepfake creation tools, thereby using Artificial Intelligence to counter Artificial Intelligence.

Outline of the Paper

This paper is structured to provide a comprehensive overview of the deepfake detection system. Following this introduction, Section 4 presents a detailed literature review of existing deepfake detection approaches. Section 5 outlines the methodology employed, including the research design, data collection, and analysis techniques. Section 6 presents the results obtained from the evaluation of our model. Section 7 discusses the implications of these results and compares them with previous work. Finally, Section 8 provides the conclusion and suggests future research directions.

3. LITERATURE REVIEW

Previous research has explored various approaches to deepfake detection, each with its strengths and limitations.

Face Warping Artifacts

One approach, “Face Warping Artifacts”, focuses on detecting artifacts by comparing generated face areas with their surrounding regions using a dedicated Convolutional Neural Network (CNN) model. This method is based on the observation that current deepfake algorithms generate images of limited resolutions, which are then transformed to match the faces in the source video. A limitation of this method is its lack of consideration for the temporal analysis of frames.

Detection by Eye Blinking

Another method, “Detection by Eye Blinking”, identifies deepfakes by analyzing eye blinking patterns as a crucial parameter for classification. The Long-term Recurrent Convolution Network (LRCN) was utilized for temporal analysis of cropped eye blinking frames. However, as deepfake generation algorithms have advanced, the absence of eye blinking is no longer a definitive clue for detection. More comprehensive parameters, such as teeth enhancements, facial wrinkles, and eyebrow placement, must be considered.

Capsule Networks for Forged Images and Videos

Research on "Capsule networks to detect forged images and videos" proposes using a capsule network to detect manipulated images and videos in various scenarios, including replay attack detection and computer-generated video detection. A drawback of their method is the use of random noise during the training phase, which, despite beneficial performance on their dataset, may lead to failure on real-time data due to noise in training. Our proposed method aims to be trained on noiseless and real-time datasets.

Recurrent Neural Networks for Deepfake Detection

The application of "Recurrent Neural Network (RNN) for deepfake detection" involves using RNN for sequential processing of frames in conjunction with an ImageNet pre-trained model. This approach utilized the HOHO dataset, which consisted of only 600 videos. The small size and homogeneous nature of this dataset may lead to suboptimal performance on real-time data. Our model will be trained on a significantly larger volume of real-time data.

Synthetic Portrait Videos using Biological Signals

The "Synthetic Portrait Videos using Biological Signals" approach extracts biological signals from facial regions in pristine and deepfake portrait video pairs. It applies transformations to compute spatial coherence and temporal consistency, captures signal characteristics in feature vectors and photoplethysmography (PPG) maps, and further trains a probabilistic Support Vector Machine (SVM) and a Convolutional Neural Network (CNN). The average of authenticity probabilities is then used for video classification. While "Fake Catcher" within this approach aims for high accuracy independent of generator, content, resolution, and video quality, a limitation is the lack of a discriminator, which complicates formulating a differentiable loss function to preserve biological signals.

4. METHODOLOGY

Research Design

Our deepfake detection system employs a deep learning-based approach, combining a Res- Next Convolutional Neural Network (CNN) with a Long Short Term Memory (LSTM) based Recurrent Neural Network (RNN). The Res-Next CNN is utilized to extract robust frame-level features from the input videos. These extracted features are then fed into the LSTM-based RNN, which is well-suited for processing sequential temporal information present in video frames. This combined architecture allows the system to analyze both spatial characteristics within frames and temporal consistency across frames, enabling a comprehensive classification of videos as either deepfake or real.

Data Collection

To ensure the model's effectiveness in real-time scenarios, it was trained and evaluated on a large, balanced, and mixed dataset. This dataset was meticulously prepared by combining various publicly available datasets, including Face-Forensic++, Deepfake detection challenge, and Celeb-DF. Additionally, real-time data from YouTube was incorporated to enhance the model's performance on diverse and realistic video content. This comprehensive data strategy ensures that the model learns features from a wide range of deepfake types and real videos, minimizing bias and improving generalization.

Sample/Population

The datasets used for training and evaluation were chosen to provide a balanced representation of both real and synthetic videos. This balanced training approach was identified through extensive research as the optimal strategy to prevent the introduction of bias and variance into the algorithm, thereby achieving higher accuracy. The selection of varied datasets ensures that the model is exposed to different deepfake generation techniques and real video characteristics, leading to a more robust and generalizable detection system.

Data Analysis

The accuracy of the trained deepfake detection model was evaluated using the Confusion Matrix approach. This method provides a clear breakdown of the model's performance, indicating true positives, true negatives, false positives, and false negatives, thereby offering a comprehensive understanding of its classification capabilities. Key parameters identified for analysis during the problem-solving methodology included: blinking of eyes, teeth enchantment, distance between eyes, presence of moustaches, double edges (eyes, ears, nose), iris segmentation, wrinkles on face, inconsistent head pose, face angle, skin tone, facial expressions, lighting, different poses, double chins, and hairstyle, and higher cheek bones.

Ethical Considerations

Ensuring the safety and integrity of user data is paramount. The system is designed to preserve data integrity, meaning that once a video is uploaded, it is processed solely by the algorithm, with strict measures to secure it from human intervention. Furthermore, to enhance the safety and privacy of uploaded videos, they are automatically deleted from the server after 30 minutes. This policy underscores our commitment to responsible data handling and user privacy in the context of deepfake detection.

5. RESULTS

The developed deepfake detection system achieved competitive results, demonstrating a simple yet robust approach to distinguishing AI-generated fake videos from real ones. The model's performance was evaluated using a large, balanced, and mixed dataset, including real-time YouTube videos. The outcome of the solution is a trained deepfake detection model designed to help users verify the authenticity of new videos. The web-based application allows users to upload videos, which are then pre-processed and classified by the model as either deepfake or real, with a confidence score provided to the user.

6. DISCUSSION

The findings from this research highlight the effectiveness of combining Res-Next CNN for feature extraction and LSTM-based RNN for temporal analysis in deepfake detection. This hybrid model successfully identifies the subtle artifacts left by deepfake generation tools, which are often imperceptible to the human eye. The comprehensive training on a diverse and balanced dataset, including Face-Forensic++, Deepfake detection challenge, Celeb-DF, and YouTube videos, has enabled the model to perform well across various deepfake types and real-world scenarios.

Compared to existing literature, our method addresses some limitations observed in previous studies. For instance, unlike methods reliant solely on eye blinking, our system considers a broader range of facial parameters, making it more resilient to advanced deepfake techniques. Furthermore, by training on noiseless and real-time datasets, we mitigate the issues encountered by capsule networks that used random noise during training, which could lead to performance degradation on real-time data. The use of a large and varied dataset also overcomes the limitations of models trained on smaller, less diverse datasets, such as the HOHO dataset.

The implications of this research are significant for combating the increasing spread of malicious deepfakes. By providing an easy-to-use web-based platform, the system democratizes access to deepfake detection capabilities, empowering individuals to verify video authenticity. The potential for integration into larger platforms like WhatsApp and Facebook could enable proactive detection and prevent the percolation of deepfakes on a wider scale.

However, the study also has limitations. While the model shows competitive results, the dynamic and evolving nature of deepfake generation techniques means continuous updates and retraining will be necessary to maintain high accuracy. The current system focuses on video analysis, and future work could explore multimodal detection incorporating audio analysis. The hardware requirements for processing, particularly for video batch processing, are substantial, suggesting a need for optimization for broader accessibility.

7. CONCLUSION AND FUTURE SCOPE

Conclusion

This research successfully developed and evaluated a deep learning-based system for deepfake detection, addressing the critical need to distinguish AI-generated fake videos from real ones. By combining Res-Next CNN for spatial feature extraction and LSTM-based RNN for temporal analysis, the system effectively identifies subtle deepfake artifacts. Comprehensive training on a diverse and balanced dataset, including real-time videos, has resulted in a robust and competitive model. The project's outcome is a functional deepfake detection model and a user-friendly web-based application that contributes significantly to mitigating the threats posed by deepfakes on the internet.

Future Scope

The future scope of this project is extensive. The web-based platform could be scaled up to a browser plugin for automatic deepfake detection. Integration with major social media applications like WhatsApp and Facebook could facilitate real-time pre-detection of deepfakes before dissemination. Further research could explore the incorporation of additional biological signals and multimodal analysis (e.g., audio analysis) to enhance detection accuracy and robustness. Continuous model retraining with new deepfake generation techniques will be crucial to maintain efficacy against evolving threats. Optimization for reduced processing power requirements would also expand its accessibility and usability across various devices.

A. REFERENCES

[1] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6.

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, June 2008. Anchorage, AK.

[3] Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin "Detection of Synthetic Portrait Videos using Biological Signals" in arXiv:1901.02212v2 .

[4] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv:1412.6980, Dec. 2014.

[5] ResNext Model: https://pytorch.org/hub/pytorch_vision_resnext/ accessed on 06 April 2020 .

[6] https://www.geeksforgeeks.org/software-engineering-cocomo-model/ Accessed on 15 April 2020

[7] Deepfake Video Detection using Neural Networks http://www.ijsrd.com/articles/IJSRDV8I10860.

[8] FaceApp: https://faceapp.com/

[9] Face Swap: https://faceswap.dev/

[10]Wikipedia: https://en.wikipedia.org/wiki/Political_tension

[11]COCOMO Model: https://www.geeksforgeeks.org/software-engineering-cocomo- model/

B. Project Planner

(As the detailed project planner in Appendix B of the thesis is not provided as text, this section remains a placeholder for the user to consult their original document. However, the thesis mentions a 12-month time estimate.)

C. Paper Published Summary

(As the paper published summary in Appendix C of the thesis is not provided as text, this section remains a placeholder for the user to consult their original document.)

D. Participation Certificate

(As the participation certificate in Appendix D of the thesis is not provided as text, this section remains a placeholder for the user to consult their original document.)

E. Plagiarism Report

(As the plagiarism report in Appendix E of the thesis is not provided as text, this section remains a placeholder for the user to consult their original document.)

F. Information of Project Group Members

(As the information of project group members in Appendix F of the thesis is not provided as text, this section remains a placeholder for the user to consult their original document.)