World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

A survey on deepfake detection through deep learning

P. Kamakshi Thai, Sathvik Kalige, Sai Nikhil Ediga * and Lokesh Chougoni

Department of CSE (Artificial Intelligence & Machine Learning), ACE Engineering College, Hyderabad, Telangana, India.

World Journal of Advanced Research and Reviews, 2024, 21(03), 2214–2217

Publication history: Received on 15 February 2024; revised on 23 March 2024; accepted on 26 March 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.21.3.0946

Abstract

Imagine watching a video where Tom Hanks delivers a rousing speech, but you suspect it might be fabricated. This growing concern stems from the rise of"DeepFakes," hyper realistic manipulated videos created using deep learning algorithms. These tools can seamlessly stitch together faces, voices, and body movements, blurring the lines between reality and fiction. While DeepFakes hold promise for entertainment and creative expression, their potential for misuse is significant. Malicious actors could leverage them to spread misinformation, damagereputations, or even influence elections. Thankfully, researchers are developing sophisticated techniques to detect these synthetic creations. This survey delves intothe realm of DeepFake detection, exploring various methods employed by deep neural networks (DNNs). We'll dissect how DeepFakes are made, categorize the most common creation techniques, and analyze the strengths and weaknesses of different detection approaches. Furthermore, we'll examine the evolving landscape of DeepFake datasets, which fuel the training and testing of detection models. We'llalso discuss the ongoing quest for a universal DeepFake detector, capable ofidentifying even unseen manipulations. Finally, we'll touch on the ongoing challenges facing both DeepFake creators and detectors, highlighting the arms racethat is unfolding in this technological battleground. By shedding light on these advancements and obstacles, we hope to empower audiences with the knowledge tocritically evaluate the information they encounter in the digital age.

Keywords: Deep learning; DeepFake; CNNs; GANs; MobileNet; Feature Extraction; Anomaly Detection; Temporal Analysis; Explainable AI (XAI)

1. Introduction

Detecting fake documents has become increasingly critical in today's digital age due to the pervasive use of digital data in various aspects of life, including digital marketing, legal forensics, medical imaging, and satellite imagery processing. This reliance on digital information has unfortunately led to a rise in cybercrime and a decline in the trustworthiness of digital data. In response to these challenges, significant efforts have been made in the field of multimedia forensics over the past 15 years. This research, driven by academic communities, major IT firms, and government organizations, aims to develop methodologies and tools for verifying the integrity of digital media. Projects like the U.S. Department of Defense's MediFor have played a crucial role in advancing research in this area by providing resources and benchmarks for evaluating media integrity.

Among the technologies that have revolutionized multimedia forensics is Convolutional Neural Networks (CNNs). Originating from the neo cognition concept proposed by Kunihiko Fukushima in 1979, CNNs have emerged as powerful tools in computer vision and robotics. Notably, LeNet-5, developed by Le-Cun et al., demonstrated significant success in handwritten digit classification, showcasing the potential of CNN architectures. CNNs are structured with convolutional, pooling, and fully connected layers, each playing a unique role in feature extraction and classification. The convolutional layer, in particular, is instrumental in identifying patterns within data by applying kernels across input tensors to generate feature maps. Through processes like forward propagation and backpropagation, CNNs learn to recognize and

* Corresponding author: Ediga Sai Nikhil

Copyright © 2024 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

classify patterns, making them indispensable in the detection of fake documents and the verification of digital information integrity. By harnessing the capabilities of CNNs and other advanced technologies, researchers continue to explore innovative solutions for addressing the challenges posed by fake documents and ensuring the reliability of digital data in an increasingly digitized world.

2. Literature review

The research centered on a thorough investigation into the effectiveness of various methodologies within the domain. By scrutinizing pertinent research papers, the aim was to assess a multitude of approaches and techniques employed in these areas. This process sought to reveal the nuanced intricacies and advancements within the field.

Xin Yang, et. al. [1] proposed a Deepfake detection system based on inconsistent head poses, observing that Deepfakes often generate mismatched facial landmarks due to face interchange. Their method employs DLib for face detection and OpenFace2 for creating a standard facial 3D model, enabling differentiation between real and Deepfake content. Using the UADFV dataset, they trained an SVM classifier with RBF kernels, achieving an AUROC of 0.89. The study emphasizes understanding Deepfake generation and highlights the potential of 3D pose estimation in detecting synthesized videos.

Rohita Jagdale, et. al. [2] have introduced a novel algorithm, NA-VSR, for Super Resolution. The algorithm begins by processing the low-resolution video into frames, followed by noise removal using a median filter. Subsequently, bicubic interpolation is employed to increase pixel density, and Bicubic transformation and image enhancement techniques are applied to enhance resolution. The design metric is evaluated using the output peak signal-to-noise ratio (PSNR) and the structural similarity index method (SSIM) to assess image quality. Results indicate significant improvements in PSNR, with the proposed method showcasing enhancements of 7.84 dB, 6.92 dB, and 7.42 dB compared to bicubic, SRCNN, and ASDS methods, respectively.

Siwei Lyu,[3] has conducted a comprehensive survey addressing challenges and research prospects in the realm of Deepfakes. Notably, a key limitation of current DeepFake generation techniques lies in their inability to produce high-quality details like skin texture and facial hair due to information loss during encoding. Lyu discusses three prominent methods: head puppetry, face swapping, and lip syncing, each serving different manipulation purposes such as mimicking behavior or altering speech. Detection strategies typically involve frame-level binary classification, categorized into inconsistencies in physical/physiological aspects, signal-level artifacts, and data-driven approaches employing Deep Neural Networks (DNNs). However, Lyu highlights limitations including dataset quality and social media manipulation.

Digvijay Yadav, et. al. [4] have provided a detailed explanation of deepfake techniques, particularly focusing on their ability to swap faces with high precision. They elucidate the workings of Generative Adversarial Neural Networks (GANs), comprising a generator network creating synthetic images and a discriminator network evaluating their authenticity. Deepfakes pose significant threats such as character defamation, spreading fake news, and undermining law enforcement efforts. Detection methods often leverage features like blinking patterns, but challenges include the need for extensive datasets, time-consuming training and swapping processes, and the similarity of faces and skin tones. Recurrent Neural Networks (RNNs), especially when combined with Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, are effective in detecting deepfake videos by analyzing frame changes. The study highlights the efficacy of Meso-4 and Mesoinception-4 architectures in achieving detection accuracies ranging from 95% to 98% on the Face2Face dataset.

Irene Amerini, et. al. [5] have introduced a system leveraging optical flow techniques to detect inter-frame dissimilarities, which are then utilized as features for CNN classifiers to learn. By comparing the optical flow fields between original and Deepfake videos, they observed that motion vectors around facial features, particularly the chin, exhibit more pronounced differences in real sequences compared to altered ones. This distinction aids neural networks in accurate learning. The study utilized the FaceForensics++ dataset, with 720 videos for training, 120 for validation, and 120 for testing. Two neural networks, VGG16 and ResNet50, were employed. For Face2Face videos, VGG achieved a detection accuracy of 81.61%, while ResNet50 achieved 75.46%. Notably, the paper's novelty lies in considering inter-frame dissimilarities, unlike other methods focusing solely on intra-frame inconsistencies, and addressing them through an optical flow-based CNN approach.

Peng Chen, et. al. [6] have introduced FSSPOTTER, a unified framework designed to simultaneously analyze spatial and temporal information within videos. The Spatial Feature Extractor (SFE) partitions videos into consecutive clips, processing each to generate frame-level features. Utilizing convolution layers from the VGG16 network with batch normalization, SFE extracts spatial features within intra-frames. Additionally, the framework employs a superpixel-wise

binary classification unit (SPBCU) to enhance feature extraction. The Temporal Feature Aggregator (TFG) utilizes a Bidirectional LSTM to identify temporal inconsistencies across frames. Finally, a fully connected layer followed by a softmax layer computes probabilities indicating whether the clip is real or fake. The methodology was evaluated using the FaceForensics++ dataset, demonstrating the framework's efficacy in detecting deepfakes.

Mohammed A. Younus, et. al. [7] conducted a comparative analysis of prominent Deepfake detection methods, encompassing techniques such as Background Comparison, Temporal Pattern Analysis, Eye Blinking, and Facial Artifacts. One method utilizes Long-Term Recurrent CNN (LRCN) to learn temporal patterns of eye blinking, employing a dataset comprising 49 interview and presentation videos along with their corresponding Deepfakes. Another technique employs a hybrid model integrating convolutional network DenseNet and gated recurrent unit cells to identify temporal discrepancies in background comparison, utilizing the FaceForensics++ dataset. Temporal Pattern Analysis utilizes Convolutional Neural Network (CNN) for feature extraction and Long Short-Term Memory (LSTM) for classification, utilizing a dataset comprising 600 videos sourced from multiple websites. Additionally, ResNet50 CNN models are employed to detect artifacts based on resolution inconsistencies, while Mesoscopic Analysis utilizes Meso-4 and MesoInception4 networks to identify Deepfakes. The study offers insights into various Deepfake detection approaches and suggests avenues for further feature exploration to enhance detection efficiency.

Shivangi Aneja et al. [8] introduced Deep Distribution Transfer (DDT), a transfer learning approach addressing zero and few-shot transfer challenges in forgery detection. Their method utilizes distribution-based loss formulation, outperforming baselines significantly in both scenarios. By employing an ImageNet-pretrained ResNet-18 neural network, DDT achieves 4.88% higher detection efficiency for zero-shot and 8.38% for few-shot transfers, broadening the scope of forgery detection.

XTao et al. [9] proposed a system emphasizing frame alignment and motion compensation using a sub-pixel motion compensation layer (SPMC) and motion compensation transformer (MCT) module within a CNN framework. Their method, validated on down sampled HD video clips, achieves superior PSNR (36.71) and SSIM (0.96) compared to SRCNN, offering insights into organizing frame inputs for improved results.

Table 1 Comparison of Deepfake Detection Techniques: Pros and Cons

Sr. No	Year	Technique/ Methodology	Pros	Cons
[1]	2019	Convolutional Neural Network (CNN)	CNNs are effective in capturing spatial features, making them suitable for image-based tasks. They excel in feature extraction from images, which is crucial for facial forgery detection.	CNNs may struggle with capturing temporal dependencies and may require additional mechanisms for handling video sequences.
[2]	2021	Long Short-Term Memory (LSTM)	LSTMs are capable of modelling temporal dependencies, making them suitable for analysing sequential data such as video frames.	LSTMs may struggle with capturing fine-grained spatial features, and their training can be computationally intensive.
[3]	2020	Optical Flow based CNN	Optical flow-based CNNs can capture motion information in videos, enhancing their ability to detect deepfake manipulations involving facial movements.	They maybe sensitive to noise and may struggle when faced with complex scenes or scenarios.
[4]	2020	Transfer Learning Based CNN Framework	Transfer learning can leverage pre-trained models on large datasets, improving generalization and performance on deepfake detection tasks.	Depending on the source domain, transfer learning may introduce biases or limitations in handling new or unseen deepfake variations.se and maystruggle when faced

Jin Yamanaka et al. [10] addressed the computational burden of single-image super-resolution systems by proposing a method that reduces deep CNN computational power by 10 to 100 times while maintaining high accuracy. With a reduced neural layer count from 30 to 11, their approach achieves significant computational efficiency without sacrificing accuracy, utilizing a dataset of 1,164 training images to demonstrate reduced space complexity and computational power.

The table above provides an analysis of different algorithms and features utilized for Deepfake detection, incorporating both Machine Learning and Deep Learning techniques. It's evident from the analysis that combining CNN with LSTM yields superior results and accuracy. Moreover, there's potential to enhance performance further by integrating the concept of Super Resolution.

3. Conclusion

In conclusion, as the prevalence of Deepfake videos rises globally, it becomes increasingly crucial to identify and detect them to prevent potential harm. Utilizing a range of Machine Learning and Deep Learning techniques alongside various features, researchers aim to accurately classify videos as authentic or manipulated. Among these methods, those employing CNN and LSTM stand out for their effectiveness in video classification. Throughout the research, diverse datasets containing both genuine and fake videos have been instrumental in refining these classification techniques. It's evident from the literature that the combination of CNN and LSTM consistently delivers superior results and accuracy in discerning Deepfake videos.

Compliance with ethical standards

Acknowledgement

We would like to thank our guide Mrs. P. Kamakshi Thai, Assistant Professor, Department of CSE (Artificial Intelligence & Machine Learning), Ace EngineeringCollege. Also, we are extremely grateful to Dr. MRS. SOPPARI KAVITHA, Headand Professor of Department of Computer Science (Artificial Intelligence and Machine Learning), Ace Engineering College for her support and invaluable time.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Xin Yang, Yeuzen Li and Siwei Lyu, EXPOSING DEEP FAKES USING INCONSISTENT HEAD POSES, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[2] Rohita Jagdale and Sanjeevani Shah, A Novel Algorithm for Video Super- Resolution, Proceedings of ICTIS 2018, Volume 1, Information and Communication Technology for Intelligent Systems (pp.533-544).

[3] Siwei Lyu ,DEEPFAKE DETECTION: CURRENT CHALLENGES AND NEXT STEPS,2020 IEEE International Conference on Multimedia & Expo Workshops(ICMEW).

[4] Digvijay Yadav, Sakina Salmani, Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network, Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019).IEEE Xplore Part Number: CFP19K34-ART; ISBN: 978-1-5386-8113-8.

[5] Irene Amerini, Leonardo Galteri, Roberto Caldelli, Alberto Del Bimbo, Deepfake Video Detection through Optical Flow based CNN, 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW).

[6] Peng Chen, Jin Liu, Tao Liang, Guangzhi Zhou, Hongchao Gao, Jiao Dai1, Jizhong Han, FSSPOTTER: SPOTTING FACESWAPPED VIDEO BY SPATIAL AND TEMPORAL CLUES, 2020 IEEE International Conference on Multimedia and Expo (ICME).

[7] Mohammed A. Younus, Taha M. Hasan, Abbreviated View of Deepfake Videos Detection Techniques, 2020 6th International Engineering Conference Sustainable Technology and Development (IEC).

[8] Shivangi Aneja, Matthias Nießner, Generalized Zero and Few-Shot Transfer for Facial Forgery Detection, arXiv:2006.11863v1 [cs.CV] 2020.

[9] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, Jiaya Jia, Detail-revealing Deep Video Super-resolution, 2017 IEEE International Conference on Computer Vision (ICCV).

[10] Jin Yamanaka, Shigesumi Kuwashima , and Takio Kurita ,Fast and Accurate Image Super Resolution by Deep CNN with Skip Connection and Network in Network, (2017 Springer).