International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

CNN and RNN using Deepfake detection

A. Sathiya Priya and T. Manisha *

Department of Information Technology, Dr.N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India.

International Journal of Science and Research Archive, 2024, 11(02), 613–618

Publication history: Received on 05 February 2024; revised on 16 March 2024; accepted on 19 March 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.11.2.0460

Abstract

Deep fake Detection is the task of detecting the fake images that have been generated using deep learning techniques. Deep fakes are created by using machine learning algorithms to manipulate or replace parts of an original video or image, such as the face of a person. The goal of deep fake detection is to identify such manipulations and distinguish them from real videos or images. Deep fake technology has emerged as a significant concern in recent years, presenting challenges in various fields, including media authenticity, privacy, and security.

Keywords: Deepfake; Deep Learning; Face Detection; CNN; Machine Learning

1. Introduction

In recent years, the proliferation of deep fake technology has raised significant concerns regarding the authenticity and trustworthiness of digital media. Deep fakes are synthetic media, typically videos or images, that are created using deep learning techniques to manipulate or replace the original content with fabricated material. These manipulated media can be incredibly realistic, making it difficult for viewers to discern between what is genuine and what is fake. The emergence of deep fake technology has revolutionized the landscape of content creation and manipulation. Deep fakes, a portmanteau of "deep learning" and "fake," refer to synthetic media generated using sophisticated artificial intelligence algorithms, often with startling realism. These manipulated videos, images, and audio recordings have the potential to deceive, manipulate, and spread misinformation at an unprecedented scale.

The implications of deep fake technology extend far beyond the realm of entertainment, infiltrating domains such as politics, journalism, and cybersecurity. As deep fakes become more accessible and convincing, the need for robust detection mechanisms has never been more urgent. Detecting deep fakes requires a multifaceted approach that combines cutting-edge technology, interdisciplinary expertise, and a nuanced understanding of the underlying algorithms and techniques employed by malicious actors.

2. Some of the existing techniques

2.1. Generative Adversarial Networks (GANs)

GANs consist of two neural networks, a generator and a discriminator, which are trained simultaneously in a competitive manner. The generator generates synthetic data (such as images or videos) from random noise, while the discriminator evaluates the authenticity of the generated data. Through adversarial training, the generator learns to produce increasingly realistic output, while the discriminator learns to distinguish between real and fake data. Variants of GANs, such as Conditional GANs (CGANs) and Progressively Growing GANs (PGGANs), have been applied to generate high-quality deep fakes with realistic facial features and expressions.

* Corresponding author: T. Manisha

2.2. Deep Neural Networks (DNNs)

Deep neural networks (DNNs) stand at the forefront of modern deep learning architectures, revolutionizing various fields such as computer vision, natural language processing, and speech recognition. These networks are composed of multiple layers of interconnected nodes, known as neurons, ironized into input, hidden, and output layers. Each neuron performs simple computations on its inputs and passes the result to neurons in the subsequent layer.

2.3. Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) play a crucial role in deep fake detection by leveraging their ability to model sequential data and temporal dependencies. In the context of deep fake detection, RNNs are employed to analyze the temporal characteristics and dynamics present in video or audio data. By processing sequential frames or audio samples, RNNs can capture subtle patterns and inconsistencies that may indicate the presence of deep fake manipulation. Similarly, in audio-based detection, RNNs can analyze the temporal patterns of speech and identify anomalies in the spectrogram or waveform indicative of synthetic audio generation techniques.

2.4. Long Short-Term Memory (LSTMs)

Long Short-Term Memory (LSTM) networks are pivotal in deep fake detection due to their proficiency in modelling long-range dependencies and capturing temporal dynamics within sequential data. In the realm of deep fake detection, LSTM networks excel at analysing sequences of frames in videos or audio samples, allowing them to discern subtle inconsistencies or artefacts indicative of synthetic manipulation. These networks are adept at learning patterns and correlations over extended time intervals, enabling them to effectively distinguish between authentic and manipulated content. In video-based deep fake detection, LSTM networks can scrutinize the temporal evolution of facial expressions, movements, and gestures, thereby detecting anomalies or irregularities introduced during the generation of synthetic videos.

3. Deepfake generation and detection

Deepfake is a technique that uses the Convolutional Neural Network (CNN) methods to generate fictitious photographs and videos. In this section, we first give an overview of the current applications and tools to create deepfake images and videos. Then, we discuss some deep learning detection techniques to overcome this issue.

3.1. Deepfake Generation

Convolutional Neural Network (CNN)is a form of deep neural network that has been commonly used to generate deep neural networks. One advantage of CNNs is that it is capable of learning from a training data set and creating a sample of data with the same features and characteristics. For example, CNNs can be used to swipe a “real” image or the video of a person with that of a “fake” one. The architecture of CNNs consists of two neural networks components: an encoder and decoder. First, the model uses the encoder to train on a large data set to create fake data. Then, the decoder is used to learn the fake data from realistic data. However, this model requires a large amount of data (images and videos) to generate realistic-looking faces. The encoder first receives random input seeds to generate a fake sample. Those fake samples are used to train the decoder. The decoder is simply a binary classifier, and it takes the real samples and fake samples as inputs and then, the decoder applies a SoftMax function to distinguish the realistic data from the fake one.

Many deepfake applications have already been around for quite a few years. FakeApp is the first method that has been used widely for deepfake creation. This FakeApp is capable of swapping faces on videos using an autoencoder-decoder pairing structure developed by a Reddit user. Similar to CNNs, FakeApp consists of the autoencoder which is used to construct latent features of the human face images and the decoder which is used to re-extract the features for the human face images. This simple technique is powerful as it is capable of producing extremely realistic fake videos that are hard for people to differentiate from the real one. VGGFace is another popular deepfake technique based on the generative adversarial network (CNN). The architecture was improved by adding two layers called adversarial loss and perceptual loss. Those layers are added to autoencoder-decoder to capture latent features of face images such as eye movements in order to produce more believable and realistic fake images.

The deepfake technique that extracts the characteristics of one image and produces another image with the same characteristics via the GAN architecture. This method applies a cycle loss function that enables them to learn the latent features. Dissimilar from FakeApp is an unsupervised method that can perform image-to-image conversion without using paired examples. In other words, the model learns the features of a collection of images from the source and target that do not need to be related to each other.

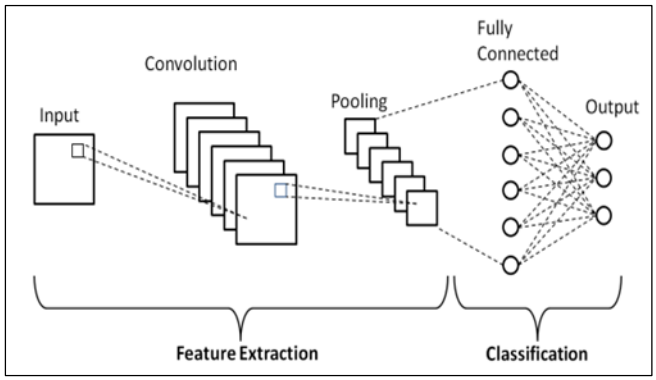

Figure 1 Classification of CNN

3.2. Deepfake Detection

Deep learning has achieved great success in deepfake detection. In this subsection below, we first discuss the Image Detection models using deep learning technologies and then Video Detection models are presented.

3.2.1. Image Detection Models

Different methods have been proposed to detect the CNN generated images using deep networks. The neural network-based methods for detecting fake CNN videos. This method employs pre-processing techniques to analyze the statistical features of image and enhances the detection of fake face image approach based on a deep convolutional neural network for detecting fake image generated by CNNs. The model first uses a deep learning network to extract face features based on face recognition networks. Then, a fine-tuning step is used to make face features suitable for real/fake image detection. These methods produce good results from the contest validation data.

However, the majority of previous research ignores the critical issue of the forensics model's generalization capabilities. In other words, they use the same type of dataset to train and test their models. To tackle this problem, it introduces a forensics convolutional neural network (CNN) that applies two image preprocessing steps to detect fake human images: Gaussian Blur and Gaussian Noise. The idea behind this model is to use preprocessing steps to neglect low level high frequency clues artifact in CNN images and improve high frequency pixel noise in low level pixel statistics. This enables the forensic classifier to learn more meaningful characteristics of real and false images, allowing it to better distinguish between real and fake image faces. The findings of the experiment reveal that the model can detect false images.

In addition to the traditional deepfake detection models, a hybrid approach was introduced to effectively detect the fake images for example proposed a two-stream network for detecting face tampering. The face classification stream is used on GoogleNet [31] to train the model on tampered and authentic images. Then, the patch triplet stream is used to analyze features using feature extractor and captures low.

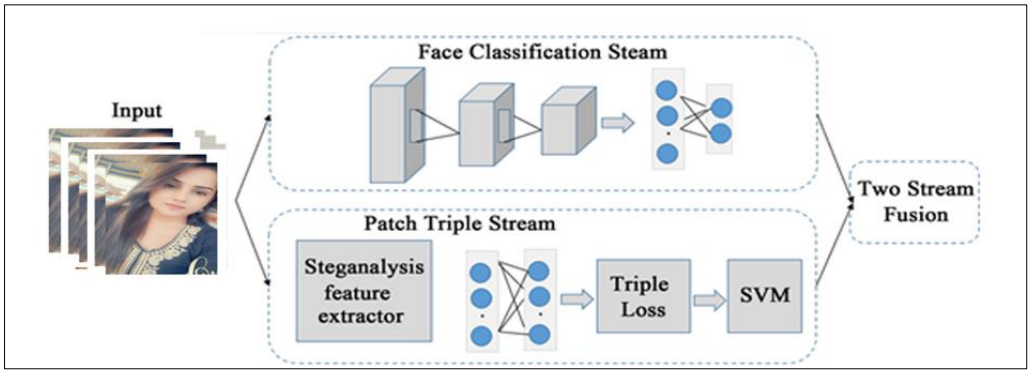

Figure 2 Two-stream neural networks

level camera characteristics and local noise residuals. The experimental results show that this approach can learn both fake and real images. Another hybrid approach was introduced which uses pairwise-learning for deepfake image detection. The approach first uses CNNs to create and generate a fake image. Then, on the popular fake feature network

(CFFN) generated by CNNs, a pairwise-learning model is used to capture the discriminant information between the fake image and the real image. The evaluation results show that this approach can overcome the shortcomings of the existing state-of-the-art fake image detectors.

3.2.2. Video Detection Models

For the last few years, deep learning methods have been successfully applied for fake image detection. However, the current deep learning methods for image cannot be directly applied for fake videos detection due to the availability of significant loss of frame information after video compression. In the subsection below, we have divided the related work in deepfake video detection into two main categories: biological singles analysis and spatial and temporal features analysis.

Biological Singles Analysis

The approach is based on a natural network to detect Fake Face Videos. Compared with the previous work, this method considers eye blinking to detect fake videos, which is an important physical feature that can be used to distinguish the fake videos. To achieve that, this method uses a convolutional neural network (CNN) with a recursive neural network (RNN) to discover the physiological signals such eye movement and blinking. Then, the model uses a binary classifier to detect the close and open eyes state. This approach is tested with a dataset called eye-blinking crawled from the internet. The eye-blinking datasets is the first available dataset which is specially designed for the eye-blinking detection. The experiment's results demonstrate the efficacy of the suggested approach in detecting false images.

Other biological signals such as heartbeat have been shown to be a reliable predictor for real video. They designed a Generative Adversarial Network (CNN) based model that can detect the deepfake video source by analyzing the "heartbeat" of deep fakes. The proposed model starts by having several detector networks where the input to this model is the real video. Then, the pair of the realistic video and fake videos is assigned to another layer called registration, which extracts facial regions of interest (ROI) and the biological signals to create PPG cells. Here, PPG cells are spatiotemporal windows which contain multiple faces extracted using a face detector. The last layer is responsible for classifying the video as fake or real. The authors used several publicly available datasets to test their model. The result shows the models have an accuracy of 97.3% in deepfake detection.

Prior research has shown that, in addition to biological signals, there is a close relationship between various audio-visual modalities of the same sample. The developed a deep learning framework for detecting deepfake in multimedia materials. The primary goal of this model is to comprehend and examine the interaction of the audio (speech) and video (visual) modalities. To achieve that, the model uses a Siamese network-based architecture to simultaneously extract the speech and face modalities. To discriminate between real and fake videos, the vector representation for the video and audio of the sample are extracted using two modality embedding networks: OpenFace and pyAudioAnalysis respectively. Finally, a triplet loss function is used to calculate the similarity and identify the fake video and the real one.

Spatial and Temporal Features Analysis

Most current deepfake detection methods only use a single video frame. In fact, video manipulation can be carried out on multiple frame-level features. Recently, many researches have shown that analyzing the temporal sequence between frames can successfully help to discriminate between the real video or the fake one. The temporally-aware model to detect deepfake videos. The model first employs a convolutional neural network (CNN) for frame features extraction. Afterwards, these features are passed to the LSTM layer to analyze a temporal sequence for face manipulation between frames. Finally, a softmax function is used to classify the video as either real or fake. For the evaluation, a collection of 600 videos was collected from multiple websites. The experimental results show the effectiveness of this model for deepfake video detection. Based on the previous version of Cycle-CNN the new approach is called Recycle-CNN, which uses conditional generative adversarial networks to merge spatial and temporal data. The evaluation results show that combining the spatial and temporal constraints can produce an effective output. Furthermore, they propose a new approach based on recurrent convolutional networks. The approach consists of two analysis stages: face processing stage followed by face manipulation detection. In the processing, face cropping and alignment is extracted using Spatial Transformer Network (STN). Then, the output from the previous stages is passed for face manipulation detection using the recurrent convolutional network, where the temporal information across frames is analyzed.

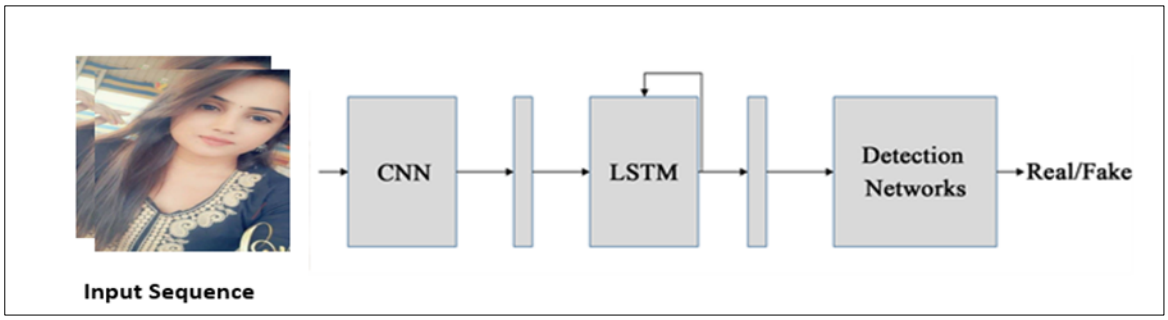

Figure 3 Convolutional neural network for spatial and temporal features analysis

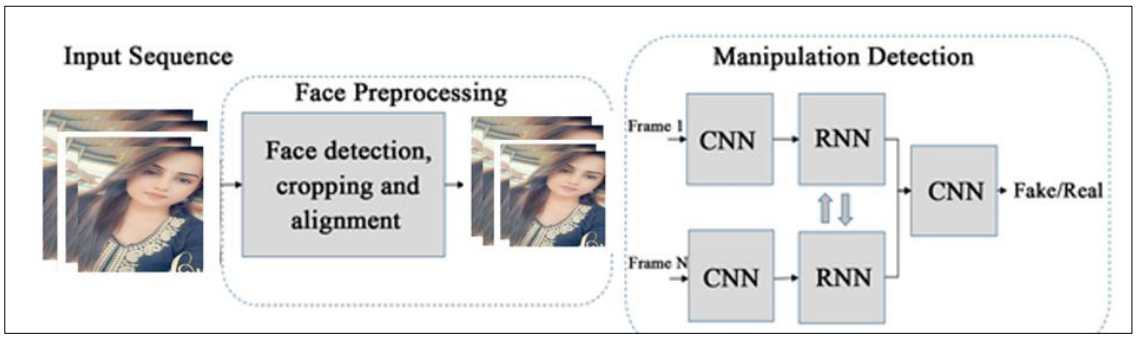

Figure 4 The proposed method is a two-step process. The first step is for face detection, cropping and alignment. The second step is for manipulation detection

3.3. Future direction and challenges

- The future direction of deepfake technology holds both promise and challenge, as advancements continue to push the boundaries of synthetic media generation while simultaneously raising concerns about its misuse and potential societal impacts. In the coming years, deepfake technology is likely to witness further refinement and sophistication, driven by advancements in machine learning, computer vision, and audio processing. These advancements may lead to the creation of even more convincing and indistinguishable deepfakes, with enhanced realism and seamless integration of visual and auditory elements. Moreover, the democratization of deepfake tools and techniques may result in their widespread accessibility, empowering individuals with the ability to generate and disseminate manipulated content on a massive scale.
- However, along with these advancements come significant challenges and ethical considerations. The proliferation of highly convincing deepfakes poses a threat to trust, authenticity, and the integrity of digital media, exacerbating existing issues related to misinformation, disinformation, and online manipulation. Deepfakes have the potential to undermine public trust in media sources, sow confusion and division, and even manipulate public opinion and elections. Moreover, the use of deepfakes for malicious purposes, such as identity theft, revenge porn, and cyberbullying, raises serious ethical and legal concerns, necessitating robust regulatory frameworks and countermeasures to protect individuals' rights and privacy.

4. Conclusion

In conclusion, while the future of deepfake technology holds immense potential for creative expression and innovation, it also presents significant challenges that must be addressed to safeguard against its misuse and mitigate its negative societal impacts. By fostering collaboration, innovation, and responsible stewardship of technology, we can harness the benefits of deepfake technology while safeguarding against its potential harms, ensuring a more trustworthy and resilient digital future.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Ahmed, S. R. A., & Sonuç, E. (2021). Deepfake detection using rationale-augmented convolutional neural networks. Applied Nanoscience, 13, 1485–1493.

[2] Fakeapp. https://www.fakeapp.org/. (Accessed on 05/29/2018).

[3] IEEE's Signal Processing Society - Camera Model Identification — Kaggle. https://www.kaggle.com/c/ sp-society-camera-model-identification/ discussion/49299. (Accessed on 05/29/2018).

[4] The Outline: Experts fear face swapping tech could start an international showdown. https://theoutline.com/post/3179/deepfake-videos-are-freaking-experts-out? zd=1&zi=hbmf4svs . (Accessed on 05/29/2018).

[5] Karen Simonyan and Andrew Zisserman, Very Deep Convolutional Networks for Large –Scale Image Recognition, ICLR 2015, arXiv:1409.1556v6 [cs.CV] 10 Apr 2015

[6] Yuezun Li, Ming-Ching Chang and SiweiLyu,In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking, arXiv:1806.02877v2 [cs.CV] 11 Jun 2018

[7] Donahue et al. Long-term recurrent convolutional networks for visual recognition and description. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4):677–691, Apr. 2017

[8] Yuezun Li, Ming-Ching Chang and SiweiLyu,In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking, arXiv:1806.02877v2 [cs.CV] 11 Jun 2018

[9] Schwartz, Oscar (12 November 2018). "You thought fake news was bad? Deep fakes are where the truth goes to die". The Guardian.

[10] A. Maclaughlin, J. Dhamala, A. Kumar, S. Venkatapathy, R. Venkatesan, and R. Gupta, Evaluating the Effectiveness of Efficient Neural Architecture Search for Sentence-Pair Tasks. .

[11] A. Hesham, Y. Omar, E. El-fakharany, and R. Fatahillah, "A Proposed Model for Fake Media Detection Using Deep Learning Techniques, in Lecture Notes on Data Engineering and Communications Technologies, vol. 152, 2023.

[12] R. M. Jasim and T. S. Atia, An evolutionary-convolutional neural network for fake image detection," Indones. J. Electr. Eng. Comput. Sci., vol. 29, no. 3, 2023, doi: 10.11591/ijeecs.v29.i3.pp1657-1667.

[13] P. Pei, X. Zhao, Y. Cao, and C. Hu, Visual Explanations for Exposing Potential Inconsistency of Deepfakes," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2023, vol. 13825 LNCS, doi: 10.1007/978-3-031-25115-3_5.

[14] C. B. Miller, Technology and the Virtue of Honesty," in Technology Ethics: A Philosophical Introduction and Readings, 2023.

[15] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: A compact facial video forgery detection network. Paper presented at the 2018 IEEE International Workshop on Information Forensics and Security (WIFS).