Global Journal of Engineering and Technology Advances

eISSN: 2582-5003

Cross Ref DOI: 10.30574/gjeta

Journal homepage: https://gjeta.com/

Global Journal of Engineering and Technology Advances
eISSN:2582-5003
Global Scholar Press, INDIA

(Research Article)

Deep fake video/image detection using deep learning

Usha MG [1, *] and Pradeep BM [2]

[1] *Department of Computer Science and Engineering, Maharaja Institute of Technology, Mysore, India.*

[2] *Assistant Professor, Department of Computer Science and Engineering, Maharaja Institute of Technology, Mysore, India.*

Global Journal of Engineering and Technology Advances, 2024, 20(02), 074–080

Publication history: Received on 16 June 2024; revised on 05 August 2024; accepted on 07 August 2024

Article DOI: https://doi.org/10.30574/gjeta.2024.20.2.0148

Abstract

With the widespread of deep fake technology, the potential to detect manipulated images has become an insistent concern. This study investigates the application of machine learning concept and its techniques, precisely CNNs (Convolutional-Neural-Networks) and LSTM (Long-Short-Term-Memory) networks to rectify deep fake images. CNNs are utilized for their strength in feature extraction from images capturing spatial hierarchies in data, while LSTMs are employed to understand the temporal dependencies that might exist in sequential frames of manipulated videos. The proposed theory combines these two architectures to harness their complementary strengths, delivering a powerful solution for detecting deep fakes. This proposed model showcases the efficiency of this hybrid approach, highlighting its potential in distinguishing between genuine and manipulated images with high accuracy. This research contributes to the development of reliable automated systems capable of mitigating the risks posed by deep fake technology. Our proposed model is trained and evaluated on a comprehensive dataset comprising both real and deep fake images with rigorous preprocessing and data augmentation techniques applied to enhance model robustness. The integration of CNN and LSTM networks leverages the strengths of both architectures enabling the model to achieve high accuracy in detecting deep fake images. Trained results demonstrate that our approach markedly improves detection rates over conventional methods, offering a dependable solution for the risks posed by deep fake technology.

Keywords: Long Short-Term Memory (LSTM); Deep Learning; Deepfake; Convolutional-Neural-Networks (CNNs); Machine Learning

1. Introduction

In the recent days, advent of deep fake technology has raised significant concerns due to its potential misuse in creating highly realistic yet fake images and videos. These synthetic media can be employed in various malicious activities, including misinformation campaigns, identity theft, and blackmail. The increasing sophistication of deep fake algorithms poses a considerable challenge for traditional detection methods, necessitating the development of advanced techniques to combat this threat.

CNNs (Convolutional-Neural-Networks) have highly effective in image detection and recognition tasks due to their potential to haphazardly extract and learn hierarchical features from digital raw pixel data. However, when dealing with sequential data, such as video frames, incorporating temporal dependencies becomes crucial.

There are 2 most important components in a LSTM those are extracting frame related information or features using CNN and sequence of data analysis using LSTM. LSTM (Long-Short-Term) networks, a type of recurrent neural-network (RNN) are suitable in capturing the image temporal data and adept at learning from sequences of data.

[*] Corresponding author: UshaMG.

This study proposes a hybrid approach integrating CNNs and LSTMs to enhance the detection of deep fake images. By exploiting the spatial feature information retrieval capabilities of CNNs and the image temporal pattern recognition strengths of LSTMs, the proposed method aims to enhance the accuracy and robustness of deep fake detection systems. Through extensive experimentation and analysis, this method and research seeks to provide a comprehensive solution to growing problem of deep fake media.

2. Literature review

Deepfakes, derived from a deep learning and fake, involve techniques which can overlay images of a men/women/person onto a video of another person, resulting in a video showing the target individual performing actions or speaking words originally done by the source person. This specific type of Deepfake is known as faceswap. Broadly, Deepfakes encompass AI-synthesized content that can also be categorized into lipsynchronization and puppet-master. Lipsynchronizationdeepfakes adjust the mouth movements in a video to align with an audio recording. Puppet-master is animating facial expression, eye and head movements of a target person to mirror those of another person captured on camera. While traditional visual effects or computer-graphics techniques can create some deep fakes, modern deep fakes predominantly rely on learning models which are autoencoders and GANs (generative adversarial networks). Creating realistic deepfakes generally requires a substantial amount of video and image data to train the models effectively. Public figures like celebrities and politicians, who have numerous data like images available online and these are the main targets for the deepfakes. This advance technology are the threat to a global security, as these can be used to fabricate videos of world leaders delivering fake speeches, potentially leading to social and religious or political tensions, misleading the public during election campaigns, or causing distress in financial market or chaos through fake or fabricated news. Furthermore, deepfakes can produce fake satellite images, such as a non-existent bridge over a river, which could mislead military analysts and troops during combat situations.

In this digital era, deepfake technologies are posing a threat to people or public confidence. Many believe advancement of this innovation would stifle people or societal growth. By superimposing another person's face over face of famous actor to target them, a deep fake image/video can be used to damage a person image and promote false or negative propaganda. And later it can lead to blackmail, shame or it can cause imbalance by disturbing peace. Hence, identifying those fraudulent images has become extremely important. The increasing sophistication of deep fake algorithms poses a considerable challenge for traditional detection methods, necessitating the development of advanced techniques to combat this threat.

Traditional methods of image authentication, which rely on manual inspection or simple digital forensics techniques, are increasingly inadequate in the face of advanced deep fake generation methods. This inadequacy necessitates the development of more robust and automated detection systems. Among various machine learning approaches, Convolutional Neural Networks (CNNs) is effective and it have proven effective in extracting complex features from images, making them a popular choice for image classification tasks. However, deep fake detection requires more than just spatial analysis; it also benefits from understanding temporal sequences and patterns, which can reveal subtle inconsistencies across frames in a video or sequences of manipulated images. In another way, typical image detection approaches are not suitable for video since image data quality becomes an issue. To address this problem, automated approaches for detecting deepfake images and videos have become highly essential and the potential impact could be on the peace and security.

To address this need, this research explores the integration of CNNs with LSTM (Long-Short-Term-Memory) networks. LSTMs, a kind of recurrent-neural-network (RNN), are adept at handling sequential image data and capturing temporal data dependencies. By combining the image spatial feature retrieval capabilities of CNNs with the sequential temporal analysis strengths of LSTMs, we aim to develop a hybrid model that enhances the accuracy and reliability of deep fake detection.

The proposed system begins with a CNN that processes each image to extract detailed spatial features. These features are then passed to an LSTM (Long-Short-Term-Memory) network, which examines the temporal relationships and the patterns within the sequence of features. This dual approach allows the model to detect inconsistencies that might be missed by using CNNs or LSTMs alone.

The importance of this research held in its potential to provide a more comprehensive solution to deep fake detection. By leveraging the combined sturdiness of CNNs and LSTMs, our model aims to improve detection rates and reduce false positives. This hybrid approach is particularly relevant in applications where the authenticity of visual content is critical, such as in legal evidence, news media, and social media platforms

2.1. Specialist

SiweiLyu is a leading researcher in digital media forensics, specializing in the detection of altered images and videos, including deepfakes. His work focuses on creating algorithms that detect visual anomalies and inconsistencies to differentiate genuine media from fabricated ones. Lyu has significantly advanced the field with his research and publications and is committed to increasing public awareness about deepfake technology.

Hany Farid's research encompasses image analysis, computer vision, and digital forensics. He is well-known for creating algorithms that detect image and video manipulations. His work has been fundamental in advancing the understanding and detection of deep fakes. Farid has published numerous papers on media authentication and frequently provides expert consultation to government and industry on digital manipulation issues.

Anderson Rocha's research interests lie in artificial intelligence and multimedia forensics. He has developed various methods for detecting tampered media, including deep fakes, by leveraging machine learning and pattern recognition techniques. Rocha's innovative approaches have been widely recognized, and his extensive publication record includes numerous influential articles in the field.

Jessica Fridrich specializes in steganography, steganalysis, and digital media forensics. Her research involves creating methods to detect hidden data within images and identifying manipulated media. Fridrich has made substantial contributions to digital forensics through her pioneering algorithms and extensive publications, establishing herself as a leading expert in the field.

Dr. Christian Riess specializes in image and video forensics, with a focus on detecting digital tampering. His research includes developing techniques to identify deep fakes and other forms of media manipulation using machine learning. Riess’s contributions to the field are significant, with several key studies that enhance the understanding and detection of manipulated media. He is an active member of the academic community, frequently contributing to conferences and journals on digital forensics and media security.

3. Proposed Methodology

The Proposed Methods for deepfake image or video detection leverages the sturdiness of CNNs for spatial feature retrieval and LSTM (Long-Short-Term-Memory) networks for image temporal sequence analysis. This combining model aims to effectively capture both the spatial inconsistencies and temporal dynamics that are indicative of deepfakes.

3.1. Data Collection and Preprocessing

Dataset Compilation: Collect a comprehensive dataset containing both genuine, real and fake images and videos from different sources. This dataset should include a wide range of subjects, environments and manipulation techniques to ensure robustness.

Data Augmentation: when training a CNN model applying data augmentation method such as rotation, cropping, flipping and scaling to intense the variety of the training data also boost the model's generalization capabilities.

Preprocessing: Normalize the images and videos to a consistent format and size. Extract frames from videos at regular intervals to create a sequence of images for temporal analysis.

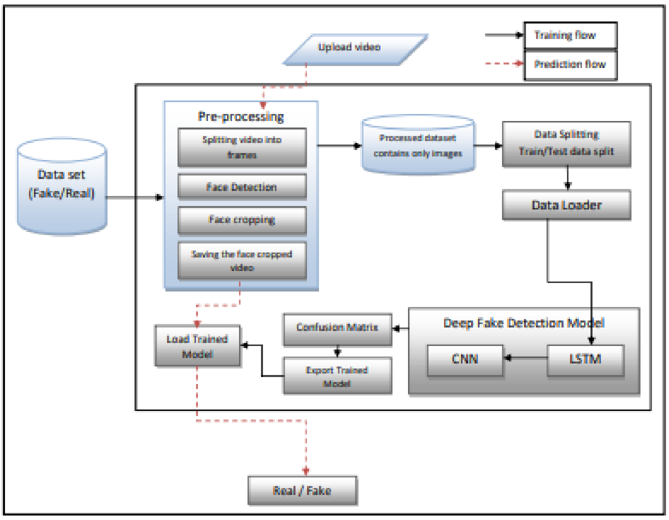

Figure 1 Proposed system architecture

3.2. Spatial Feature Extraction Using CNNs

CNN Architecture: Design CNN architecture tailored for feature extraction from images. Common architectures such as VGG16, ResNet or Inception can be used as a base with modifications to suit the specific requirements of deep fake detection.

Training the CNN: Train the CNN on the preprocessed dataset to learn spatial features that differentiate between authentic and manipulated images. Use a large labelled dataset and apply techniques such as transfer learning if necessary.

Feature Maps: Extract high-level feature maps from the CNN's intermediate layers. These feature maps represent spatial information that will be used in the subsequent temporal analysis.

3.3. Temporal Analysis Using LSTMs

Sequence Formation: Arrange the extracted feature maps from the CNN into sequences corresponding to the frames of videos. Each sequence represents the temporal evolution of features in a video.

LSTM Architecture: Design an LSTM network to process the sequences of feature maps. The LSTM will learn to recognize temporal patterns and dependencies that are characteristic of deepfake videos.

Training the LSTM: Train the LSTM on sequences of feature maps to learn temporal dynamics. Use appropriate loss functions and regularization techniques to prevent over fitting and improve generalization.

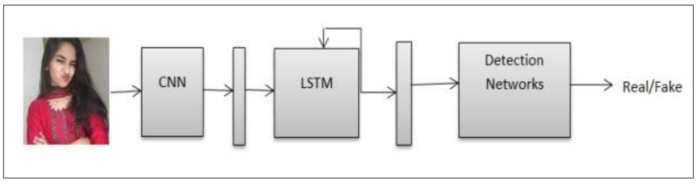

Figure 2 CNN for spatial and temporal feature analysis

4. Algorithm design

Programming Languages and Frameworks

Python: The primary programming language used for developing machine learning models due to its extensive libraries and ease of use.

TensorFlow: A powerful open-source end to end machine learning framework created by Google for constructing and training neural networks.

Keras: Keras is an open-source library that offers a Python interface. It is an advanced neural networks API, built in Python, that can run on top of TensorFlow, making model development and experimentation more efficient

4.1. Long-Short Term Memory

LSTMs (Long-Short-Term-Memory-networks) are sophisticated or advance type of RNN (recurrent-neural-network) designed for deep learning tasks that involve sequential data. LSTMs mitigate the vanishing gradient issue commonly found in traditional RNNs, enabling them to successfully capture long-term dependencies. The primary components of an LSTM consists of forget input and output gate. The forget gate determines whether information from the previous time step should be kept or discarded, the input gate processes new information from the current input, and the output gate transfers the updated information to the next time step. This architecture enables LSTMs to effectively capture and preserve crucial information across long sequences. Each iteration of these processes within an LSTM represents a single time step.

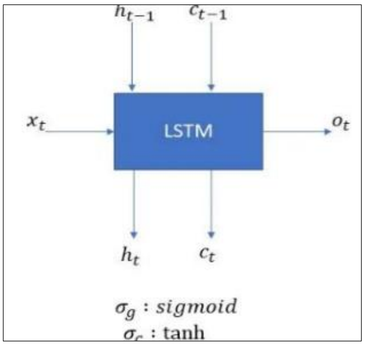

Figure 3: Multiplication of elements

$$f_t(Forget\ gate) = \sigma_g\,(W_f \times x_t + U_f \times h_{t-1} + b_f)$$
$$i_t(Input\ gate) = \sigma_g\,(W_i \times x_t + U_i \times h_{t-1} + b_i)$$
$$o_t(Output\ gate) = \sigma_g\,(W_o \times x_t + U_o \times h_{t-1} + b_o)$$
$$c'_t = \sigma_g\,(W_c \times x_t + U_c \times h_{t-1} + b_c)$$
$$c_t = f_t . c_{t-1} + i_t . c'_t$$
$$h_t = o_t . \sigma_c(c_t)$$

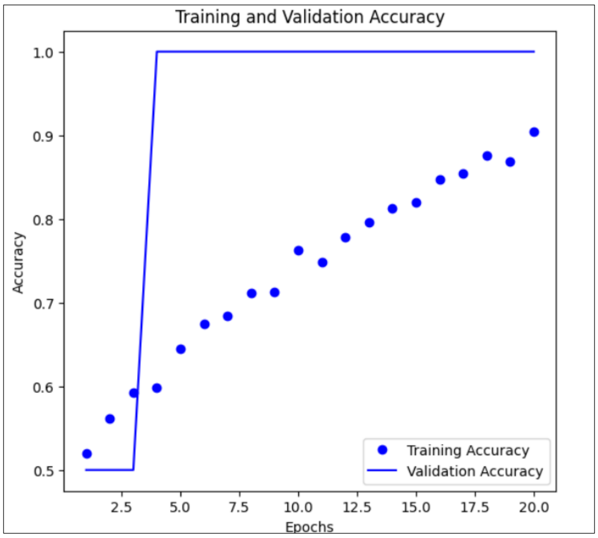

Figure 4 Accuracy of the trained model

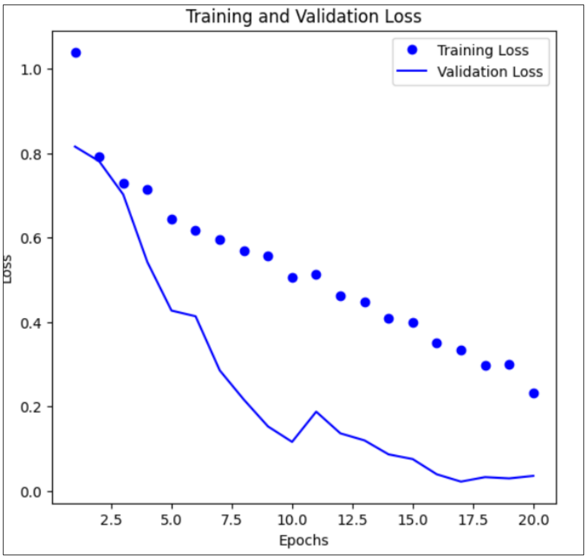

Figure 5 Loss of the trained model

Fig 3.1 and 3.2 reflects the performance of a machine learning model during its 20th epoch of training. The model processed all 65 batches with a training loss of 0.203, indicating a fairly good fit to the training data, and achieved a training accuracy of 90.45%. Despite this, the validation loss for this epoch was 0.0354, which, although low, did not surpass the best validation loss recorded so far (0.02178). However, the model showed perfect accuracy (100%) on the validation set, demonstrating strong performance on the unseen data.

4.2. Development Environments

Jupyter Notebook: An open source interactive web application that can be used for easy development and visualization, sharing of code, ideal for experimentation and iterative development.

Google Colab: A free cloud service that provides Jupyter Notebooks with free access to GPUs, making it convenient for developing and testing models without local computational constraints.

4.3. Data Handling and Preprocessing

- Pandas: A data manipulation and analysis library, useful for handling large datasets and training and preparing data.
- NumPy: A python library that performs array and matrix operating data. It is effective to handle large amount of data.
- OpenCV: An free open-source for the computer-vision library designed to address computer vision problems that suitable for the processing of image and video.

4.4. Visualization

Matplotlib: A library for creating graphs and animated and interactive and static visualizations in Python.

Seaborn: A library developed on top of Matplotlib providing API and interface for informative statistical graphics and impressive drawing.

5. Conclusions

We have proposed a hybrid deep fake detection methodology that integrates CNNs and LSTM networks to leverage their respective strengths in spatial feature extraction and temporal sequence analysis. This approach addresses the limitations of traditional detection methods by effectively capturing both the spatial inconsistencies and temporal dynamics characteristic of deep fakes.

Our methodology begins with comprehensive data preprocessing and collection, ensuring a varying and representative dataset of both real and manipulated media. The CNN component is designed to extract high-level spatial features from images, which are crucial for identifying subtle visual artifacts introduced during manipulation. These feature maps are then processed by the LSTM component which analyzes the temporal sequences to detect inconsistencies across frames in videos.

The integrated hybrid model undergoes rigorous training and evaluation demonstrating significant improvements in detection accuracy over models that rely solely on spatial or temporal analysis. Performance scale such as recall, precision, accuracy and F1-score validate the effectiveness of our approach with cross-validation further confirming the model's resilience and ability to generalize.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] ThanhThi Nguyen, Quoc Viet Hung Nguyen, Cuong M. Nguyen, Dung Nguyen, DucThanh Nguyen, SaeidNahavandi, Fellow, IEEE, 2021. "Deep Learning for Deepfakes Creation and Detection: A Survey". IEEE Transactions on Pattern Analysis and Machine Intelligence.

[2] Gaojian Wang, Qian Jiang, Xin Jin, Xiaohui Cui, "FFR FD: Effective and Fast Detection of DeepFakes Based on Feature Point Defects", 2020.

[3] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: ICASSP 2019- 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 8261–8265.

[4] Y. Li, M.-C. Chang, S. Lyu, Inictu oculi: Exposing ai created fake videos by detecting eye blinking, in: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2018, pp. 1–7.

[5] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, H. Li, Protecting world leaders against deep fakes., in: CVPR Workshops, 2019, pp. 38–45.

[6] U. A. Ciftci, I. Demir, L. Yin, Fakecatcher: Detection of synthetic portrait videos using biological signals, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[7] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, J. Zhao, Deeprhythm: exposing deepfakes with attentional visual heartbeat rhythms, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4318–4327

[8] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, Emotions don't lie: An audiovisual deepfake detection method using affective cues, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2823