International Journal of Research and Review Techniques (IJRRT), ISSN: 3006-1075
Volume 4, Issue 2, April-June, 2025, Available online at: https://ijrrt.com

Deep Learning Approach for Robust Deep Fake Detection using CNN-GRU Architecture

Ravinder Kumar[1], Madhu Rani[2]

[1]Assistant Professor, Department of Computer Science Engineering, Rattan Institute of Technology and Management, Haryana, India
[2]Research Scholar, Department of Computer Science Engineering, Rattan Institute of Technology and Management, Haryana, India

ABSTRACT

In recent years, the rapid advancement of deepfake generation techniques has posed significant challenges for content authenticity and security. To address these concerns, the adoption of deep learning methodologies for deepfake detection has gained substantial traction due to their superior accuracy over traditional methods. Among the popular architectures, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), particularly Gated Recurrent Units (GRU), have demonstrated notable effectiveness. This study presents a CNN-GRU-based model tailored for deepfake detection, incorporating a typical pipeline that includes data preprocessing, feature extraction, and classification. The model capitalizes on subtle artifacts introduced by GAN-based deepfake generators, enabling robust detection despite the increasing realism of synthetic content. Experimental results reveal a high training accuracy of 98.93% over 10 epochs and a test accuracy of 81.97%, indicating strong generalization capabilities. Despite minor fluctuations, the validation accuracy shows an overall upward trend, affirming the model's learning efficiency and stability across training phases. The proposed CNN-GRU framework thus offers a promising and contemporary solution for detecting deepfakes in dynamic and evolving data environments.

INTRODUCTION

Today, generative artificial intelligence technologies have reached such a level of sophistication that they allow the creation of images with such advanced realism that it is increasingly difficult for the human eye to differentiate them from the originals [1]. This technological advancement represents a considerable risk to society, as it facilitates identity theft and the spread of misinformation, commonly known as "Synthetic Images" [2] or “Fake Images” [3] whose purpose is to manipulate people. A central element of this problem is "fake images," techniques that allow faces in images to be altered or replaced in extremely convincing ways. This has increased their use in fraudulent activities and digital deception, calling into question trust in the authenticity of visual content circulating online.

The proliferation of manipulated images, such as fake images, represents a significant challenge for cyber security, facilitating both impersonation and the spread of disinformation. To address this problem, it is essential to implement effective solutions for the detection and analysis of manipulated content [4].

PROBLEM STATEMENT

Sight is one of the senses we rely on most to understand reality. In some ways, we believe more in what we see with our own eyes than in what we are told or read. But what happens when we are no longer certain that what we see is real? "Synthetic Images" [2] or “Fake Images” [3]are image falsifications created by artificial intelligence that raise the terrible question of whether we should believe everything we see.

The impact of fake images goes beyond misinformation, affecting both personal and social levels by damaging reputations with fake situations, distorting public perceptions during life time and on various events and issuances, and even inciting violence with manipulated scenes. To address these challenges, the scheme is proposed to detect either the image/video is “real” or “fake” using deep learning models namely MTCNN, CNN and GRU.

JUSTIFICATION

Immersion in a vast visual culture, saturated with audiovisual products with varying degrees of fidelity to reality, has altered viewers' perceptions of what is plausible and what is authentic. Individuals face the constant challenge of distinguishing between the real and the artificial, and of discerning where manipulation begins and reality ends. This transformation has generated growing concern about the authenticity of the information consumed, since the ability to create misleading visual content has surpassed traditional verification and analysis capabilities.

With the advancement of editing and image generation technologies, the line between reality and fiction has blurred, demanding greater vigilance and sophisticated methods to guarantee the integrity of the content. The choice to research and develop a technique using machine/deep learning for the detection of deep fakes in images/videos responds to an urgent need in the field of cyber security and digital forensics.

Deep fakes have the ability to undermine trust, distort information, and affect critical areas such as politics, social media, and the economy, so it is crucial to understand the nature of these falsified contents, their potential impact, and develop effective methods to identify them and reduce their harmful effects [3].

OBJECTIVES

1. Develop a model based on deep learning techniques, using MTCNN, CNN and GRU to detect fake and real in images/videos, with the aim of improving cyber security and preventing impersonation attempts in virtual environments.
2. Collect a diverse dataset of images that will be used to train the fake or real image/videos detection model.
3. Deploy deep learning models with Tensor Flow and Keras to accurately detect fake or real images/videos.
4. Evaluate the effectiveness of the fake or real image/videos detection model by testing it with specific dataset, validating the system's accuracy in different deep learning models like MTCNN, CNN and GRU.

LITERATURE REVIEW

Due to the serious negative impact of deep fakes, all sectors of society have begun to take corresponding protective measures. In order to prevent fake videos targeting politicians from affecting national security, governments have begun to promote the formulation of standards and laws for related industries. At the same time, Internet companies such as YouTube and TikTok have also begun to regulate deep fake videos and have held many fake video detection competitions. In academia, researchers have proposed a large number of detection technologies suitable for various scenarios as technical governance measures for deep fakes. In response to the fake and detection technologies that have emerged in recent years, this article describes some of the representative technologies. Compared with other existing reviews, it more systematically considers deep fakes and detection technologies of different modal information, and also introduces adversarial attack methods for deep fake generation and detection models.

WHAT ARE DEEPFAKES?

We can describe a deepfake as the manipulation of a video, image, or audio to make it appear real without being so, using artificial intelligence. The two words used come from "fake" and "deep learning," an area of artificial intelligence that uses neural networks to simulate the behavior of a human brain. The fact that new videos, images, or audio are generated through neural networks makes it more difficult to detect potential anomalies that could corroborate their falsity. An example of a deepfake was the video of former United States President Barack Obama, made in 2018, which mixed the images and audio of an actor, but with the former president's voice and face. The use of these deception techniques can lead the population to believe it at first, but after suffering the consequences (possible fraud by believing videos of people with a good reputation that have been manipulated, blackmail by showing videos or audios that the person has not made, discrediting the person, etc.), and with greater knowledge that these techniques exist, it will no longer be known if what we see in any media is true or not, causing a lack of trust on a global level, something really serious for society [12].

A deepfake often creates an unsettling sensation, revealing subtle imperfections that betray its lack of authenticity. Training the human eye to detect these inconsistencies—such as unnatural facial movements, uneven lighting, or audio-video mismatches—and combining this skill with reliable verification techniques makes it possible to distinguish deep fakes from genuine content [12].

A deep fake often creates an unsettling sensation, revealing subtle imperfections that betray its lack of authenticity. Training the human eye to detect these inconsistencies such as unnatural facial movements, uneven lighting, or audio-video mismatches and combining this skill with reliable verification techniques makes it possible to distinguish deep fakes from genuine content [12].

COMPUTER VISION TECHNIQUES

Computer vision techniques have made great progress in the last 10 years for the present verification produced mainly by the magnitude of their tentative use: novel forms of human-computer interaction, sign language recognition, deep fake detection, facial recognition, biometrics and security, tracking the trajectory of human beings or groups; to mention some of the most important. Furthermore, innovations in CV using computers and machine learning have found direct application in this type of problems, emerging a variety of techniques and diverse approaches in their solution. However, due to the complexity of the problem, it is difficult to provide a solution. One of the problems that hinders the advancement of these techniques is that their performance is often not directly verifiable, making it difficult to assess the advantage of one algorithm or specific technique over another and the situation in which the disparity is expressed. Because of this, it is important to systematize the characterization of performance, making experimental conditions clear or developing tests using standardized problem sets accessible to the entire population, ensuring reproducibility [14-16].

- **Bitmap Digital Images**

These images are made up of a rectangular grid of pixels. An image is represented by assigning and storing color information for each pixel. This color assignment is defined by a value that varies depending on the color model applied to the digital image. Bitmap images differ from vector images in that they cannot be scaled without affecting the image's appearance [14-16].

- **Vector Digital Images**

Vector figures are chosen from mathematically derived geometric materials and are given the name vector. A vector is composed of a sequence of points with handles that oversee the structure of the line, which results from joining two of those points. When these points are connected by a curve, they give rise to the essential elements of a Bézier curve (anchor points or nodes and control handles), which model the shape of the line until the desired contour is achieved. By allowing many artistic possibilities due to their manageability, Bézier curves are the most convenient way to work in graphic design through a computer, not only for logo design but also for the creation of illustrations in general. The main advantages of these images are their light weight and the ability to scale them without losing the graphic quality of their paths or fills [14-16].

ARTIFICIAL INTELLIGENCE (AI)

AI is a field of computer science that, through the design of different algorithms, seeks to make machines adopt human-like behaviors, such as learning or decision-making. This allows for solving problems that require high performance to process large amounts of data, while also improving the way they process it. The first idea related to the systematization of logical processes appeared in 1854 when the British mathematician George Boole argued that logical behavior can be expressed mathematically in the same way that a system of equations is solved [16-19].

Almost a century later, in 1950, Alan Turing published the essay computing machinery and intelligencein which he raised the idea that a machine could have (or not) intelligence or could behave like a human being. Furthermore, Turing developed a technique to determine whether a machine had the necessary qualities to be considered intelligent. This technique, known as the Turing Test or Imitation Test, consists of a natural language recognition test [16-19].

MACHINE LEARNING (ML)

ML is one of the branches of computer science and, specifically, one of the branches of artificial intelligence. This term was coined in 1959 by American computer scientist Arthur Samuel and refers to techniques focused on developing algorithms that enable machines to learn. A machine is considered to have learned when, based on its experience, it is performing better at the task assigned to it. This allows problems to be solved without the need for explicit programming, as the machine can develop its own logic using the data provided [20-23].

The limitations of machine learning algorithms were, on the one hand, the amount of data they could use for development and, on the other, the compilation time they required. Thus, in the 2010s, thanks to technological advances in both algorithm development and computer hardware, more complex problems could be posed, and the concept of deep learning was born. Deep learning (DL)it is a branch of machine learning focused on creating more complex systems, such as ANN. These networks have a large number of neurons and seek to imitate the behavior of the human brain by analyzing large amounts of data [20-23].

PROPOSED METHODOLOGY

3.1 BACKGROUND OF EXISTING DEEP FAKE DETECTION TECHNOLOGIES

Most deep fake recognition models are supported by CNN structures. One is to detect the forged information of a single frame, such as the difference in resolution between the face area and the background, the generated face details, and the difference in distribution between the tampered image and the real image; the other method is to use the temporal attribute to detect the unnaturalness between video frames.

3.2 PROPOSED MODEL

Fake face videos have inconsistencies between frames because existing fake video methods often decompose the video into single frames and process them separately, rarely considering the correlation between previous and next frames. For example, many fake techniques use facial key point detection to locate the position of the face, but the facial key point detection technology itself has an error of several pixels. If the error is not taken into account, there will be obvious visual differences in the position of the face in the generated adjacent video frames. LSTM with GRU or optical flow estimation can model the correlation between frames. Below is detailed model for ready reference.

DATASET

As the source of model knowledge, a good dataset is crucial to model performance. Similarly, in the deep fake detection task, the dataset has also been continuously developed to make up for past deficiencies. The real videos and images of the dataset were collected mainly from the Deepfake Detection Challenge (DFDC) [54].

FACE DETECTION

Face detection is one of the first steps in facial recognition systems, as it allows us to determine which face will be processed. The goal is to determine whether a face exists in the image presented, and if so, to delineate its characteristic areas and remove the background, which will store irrelevant information for subsequent processes. According to [55], facial detection methods are divided into four categories:

- Knowledge-based methods: These are based on human knowledge of facial anthropometry.
- Methods based on invariant features: Those that do not change due to existing changes in the environment (lighting, position or location of the camera).
- Pattern-based methods: Relationship between an input image and a previously defined pattern, where the objective will be to capture main features.
- Appearance-based methods: They use models obtained through image training, taking the image as a feature vector.

RNN AND LSTM

RNN are a class of networks specialized in analyzing text data. The main characteristic of this type of network lies in it sability to model temporal relationships between elements of the sequence through an internal state of the network or hidden state, which we can consider as a memory of what the network has seen up to that moment. In this architecture, a recurring formula is functional to an input succession so that, in each given step, it depends on the new input value x and theprevious internal state h.

GATED RECURRENT UNITS (GRUS)

Like LSTMs, GRUs are also designed to handle long-term dependencies in sequential data. However, GRUs have a simpler structure and use fewer internal gates. This makes them more efficient in terms of computation and training. The procedure of a GRU cell is similar to that of an LSTM, but with fewer components. The steps of a GRU are described below: The update gate decides how much new information should be added to the hidden state. It is calculated using a sigmoid function and controls the mixing between the current input and the previous state.

IMPLEMENTATION AND RESULTS

4.1 IMPLEMENTATION

The process used to recognize deep fake images and videos comprises numerous phases. First phase starts with collecting deep fake videos/images and normal images/videos which is utilized for testing and training. At the pre-processing phase, resizing and cropping is used thereafter in feature extraction. Histogram evaluation filtering is used to perform the image pixel filtering process used in processing deep fake videos and images to avoid excessive contrast enhancement.

The next stage is feature extraction using CNN, pertaining the transfer learning model namely GRU in corporating segmentation, which is responsible to determine the timeline process using frame by frame images. Thereafter, fully connected layer enhancement with hyper-parameters optimization evolved for the classification and identification deep fake images or videos. These procedures or phases will outcome is asunder.

TRAINING AND TESTING

The data set is divided into three distinct sets: training (80%), validation (30%) and testing (20%). The training set is used to adjust the model parameters, the validation set is used to adjust the hyper parameters and perform model selection, and the test set is used to evaluate the final performance of the model. This division is essential to evaluate model performance objectively and avoid over fitting.

OPTIMIZATION USING GRU

GRU-based Sequence Model (with Masking) This Keras-based neural network is designed for handling sequence data with variable lengths using GRU (Gated Recurrent Unit) layers. The model takes two inputs: frame features input – a sequence of feature vectors. Mask input – a boolean mask indicating valid time steps (used to handle padded sequences).

The architecture begins with a GRU layer that returns sequences, enabling temporal processing of input data while respecting the mask. It is followed by another GRU layer to condense the sequence into a fixed-length representation. A dropout layer is used to prevent over fitting.

The values included in the training and validation have been extracted from the training and test accuracy from the numbers of epoch of each of the networks after having subjected them to the training and validation processes. In reference to the results below the identifier “test”, these are acquired immediately after evaluating each of the models with the images of the test set with the model.evaluate() function.

CONCLUSION AND FUTURE SCOPE

CONCLUSION

In the past few years, there has been a significant development in both the creation and detection of deep fakes. There has been a great improvement in the use of deep learning techniques for deep fake detection due to the accuracy of the results compared to non-deep learning methods. Deep neural network architectures such as CNN and RNN are widely used in the implementation of deep fake detectors. A common deep fake detection pipeline consists of a data preprocessing module, a CNN-based feature extractor, and a classification module. In addition, Deep fake detection has a great dependence on the traces left by the Deep fake generator on the Deep fake. Since the current GAN-based Deep fake generators are able to synthesize more realistic Deep fakes with minimal inconsistencies. Consequently the proposed CNN-GRU model exhibits strong performance, achieving a high training accuracy of 98.93% over 10 epochs. The validation accuracy, although displaying dynamic fluctuations, follows a generally increasing trend, indicating the model’s ability to learn effectively during training. A test accuracy of 81.97% further demonstrates the model's capability to generalize well on unseen data. The variation observed in the validation results reflects the model's sensitivity to epoch-wise changes but remains within acceptable limits. This behavior suggests that while the model is adaptive, it is also stable across different training phases. Overall, the combined CNN-GRU approach proves to be both reliable and efficient for classification tasks involving sequential data and can be considered as modern and comprehensive methods to combat Deep fake.

REFERENCES

[1] Mon Rapp, Chiara Di Lodovico, Federico Torrielli, and Luigi Di Caro. 2025. How do people experience the images created by generative artificial intelligence? An exploration of people's perceptions, appraisals, and emotions related to a Gen-AI text-to-image model and its creations. Int. J. Hum.-Comput. Stud. 193, C (Jan 2025). https://doi.org/10.1016/j.ijhcs.2024.103375

[2] Man, K., & Chahl, J. (2022). A Review of Synthetic Image Data and Its Use in Computer Vision. Journal of Imaging, 8(11), 310. https://doi.org/10.3390/jimaging8110310

[3] Sharma, D. K., Singh, B., Agarwal, S., Garg, L., Kim, C., & Jung, K.-H. (2023). A Survey of Detection and Mitigation for Fake Images on Social Media Platforms. Applied Sciences, 13(19), 10980. https://doi.org/10.3390/app131910980

[4] Moore, M., PhD. (2025, January 23). Top cyber security threats to watch in 2025. University of San Diego Online Degrees. https://onlinedegrees.sandiego.edu/top-cyber-security-threats/

[5] Lu, Jiaqi. (2023). Convolutional neural network and vision transformer for image classification. Applied and Computational Engineering. 5. 104-108. 10.54254/2755-2721/5/20230542.

[6] Ding, Xinyi & Raziei, Zohreh & Larson, Eric & Olinick, Eli & Krueger, Paul & Hahsler, Michael. (2020). Swapped face detection using deep learning and subjective assessment. EURASIP Journal on Information Security. 2020. 10.1186/s13635-020-00109-8.

[7] van der Sloot, B., & Wagensveld, Y. (2022). Deepfakes: Regulatory challenges for the synthetic society. Computer Law & Security Review, 46, 1-15. Article 105716. https://doi.org/10.1016/j.clsr.2022.105716

[8] 7 examples of social media scams you should avoid at all costs. (n.d.). https://www.terranovasecurity.com/blog/examples-social-media-scams

[9] The use of AI deepfakes in cyberbullying. (n.d.). https://www.linewize.com/blog/ai-deepfakes-cyberbullying

[10] DJangi, P. (2024, July 24). These manipulated photos are the original political deepfakes. History. https://www.nationalgeographic.com/history/article/political-photo-manipulation-in-history

[11] Ghilom, M., & Latifi, S. (2024). The Role of Machine Learning in Advanced Biometric Systems. Electronics, 13(13), 2667. https://doi.org/10.3390/electronics13132667

[12] Westerlund, M. (2021). Deepfakes: A realistic assessment of potentials, risks, and policy regulation. Springer. https://doi.org/10.1007/978-3-030-63107-0

[13] Lanham, M. (2021). Generating a new reality: From autoencoders and adversarial networks to deepfakes. Apress. https://doi.org/10.1007/978-1-4842-7092-9

[14] Szeliski, R. (2022). Computer vision: Algorithms and applications (2nd ed.). Springer. https://doi.org/10.1007/978-3-030-34372-0

[15] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.