JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Deepfake Detection

Dr. P.Sruthi [1], Dr.T.Bhaskar[2], Bommagalla Ankitha[3], Duggirala Mercy Sunada[4], Yellanur Bhargavi[5] , Yerra Manasa[6]
[1]Associate Professor, Department of CSE (AI&ML), CMR College of Engineering & Technology, Hyderabad,Telangana State, India.
[2]Associate Professor, Department of CSE (AI&ML), CMR College of Engineering & Technology, Hyderabad, Telangana State, India.

***Abstract*—We provide a novel method for detecting synthetic face modifications in photographs using deep learning techniques in light of the growing threat posed by deepfake images. Our approach is designed to tackle problems caused by AI-generated deepfakes; specifically, it focuses on distinguishing real from fake facial images. A ResNetV1.5 Convolutional Neural Network (CNN), which is skilled at extracting complex features from facial data, is at the heart of our system. These characteristics are then used to train a recurrent neural network (RNN) of the Long Short-Term Memory (LSTM) type for image-level classification, which allows it to discern between real and fraudulent information.**

In order to verify the stability and practicality of our model, we carry out thorough tests with the VGGFace2 dataset, which is a well-known tool for face recognition studies. Our approach involves combining multiple datasets, such as FaceForensic++, Deepfake Detection Challenge, and Celeb-DF, to create a balanced and diversified dataset that is representative of real-world situations. Moreover, our approach prioritizes efficiency and simplicity, enabling competitive performance in the detection of altered imagery.

Our study uses AI technology to tackle the urgent problem of the spread of deepfakes. By combining an LSTM-based RNN with a ResNetV1.5 CNN, we are able to identify fake face modifications with impressive accuracy. Our thorough analysis of a variety of datasets highlights the effectiveness and usefulness of our strategy in halting the spread of deepfake images in practical contexts.

Keywords: Synthetic face modification, CNN, LSTM, Deep learning, MTCNN

1. INTRODUCTION

The emergence of deepfake images is a serious danger to the integrity of digital media, since it erodes public confidence and aids in the dissemination of misleading information. Researchers are creating advanced deep learning methods, such as recurrent neural networks with long short-term memory (LSTM) and convolutional neural networks (CNNs), to counter this threat. These architectures are very good at extracting fine-grained information from images, which makes it possible to identify the subtle changes seen in deepfakes. CNNs, which are well-known for their capacity to detect patterns and details, are an essential tool for detecting anomalies in images. On the other hand, LSTM networks supplement this power by detecting temporal dynamics in deepfake films, which gradually improves detection accuracy.

Furthermore, deepfake detection capabilities are further strengthened by the incorporation of sophisticated techniques like the Multi-Task Cascading Neural Network (MTCNN) and large-scale datasets like VGGFace2. These resources offer a wealth of training data so that models can reliably discriminate between changed and authentic faces. Furthermore, interpretability and transparency are improved by means of tools such as GradCAM visualization, which clarifies the deep learning models' decision-making process by emphasizing important areas in images. Researchers hope to strengthen defenses against deepfake manipulation, protect the integrity of visual content in the digital age, and lessen the hazards connected with it for society by utilizing these state-of-the-art technologies and approaches.

1.1 Robust Long Short-Term Memory (LSTM) neural networks:

Recurrent neural networks with Long Short-Term Memory (LSTM) are essential for identifying temporal anomalies in deepfake images. LSTM networks are particularly good at capturing these minor anomalies over time, since deepfake manipulation frequently entails changing the temporal dynamics of visual material, such as eyes, nose, ears. LSTMs contribute to the reliable identification of deepfake images by identifying patterns suggestive of manipulation by evaluating sequential data, such as frames in an animation.

.

1.2 CNN:

Deepfake images can be used to extract intrinsic characteristics thanks to Convolutional Neural Networks (CNNs), especially ResNetV1.5. Convolutional filters are used to input images by CNNs, which allows them to recognize complex patterns and details found in both real and artificial images. CNNs contribute to accurate identification by detecting tiny inconsistencies suggestive of deepfake manipulation by examining pixel-level information.

1.3 The VGGFace2 Dataset:

To train deep learning models for deepfake picture recognition, the VGGFace2 dataset is a useful tool. VGGFace2, a vast set of face photos with a range of identities and emotions, allows CNNs to be robustly trained

for facial recognition applications. Our deep learning models get a thorough comprehension of face traits and expressions, enabling them to detect deepfake manipulation with precision by utilizing the rich data provided by VGGFace2.

1.4 MTCNN:

To enhance the efficacy of deepfake recognition of pictures, the Multi-Task Cascade Neural Network (MTCNN) enables comprehensive evaluation of facial images. MTCNN improves the precision and effectiveness of the detection process by cascading activities including face detection, alignment, and feature extraction. We guarantee comprehensive analysis of face photos by incorporating MTCNN into our deep learning pipeline, which allows for reliable identification of deepfake manipulation.

1.5 Visualization with GradCAM:

The GradCAM visualization offers important insights into how our deep learning models for deepfake image identification make decisions. GradCAM provides transparency and interpretability by producing heatmaps that highlight important locations in the images that contribute to categorization results. This improves the detection system's dependability and credibility by allowing researchers to comprehend how the models recognize and distinguish between real and altered photos.

2. RELATED WORK

The ability to identify deepfake photos is becoming more and more important in the fight against the proliferation of modified visual information on different web platforms. The development of advanced detection methods to differentiate between actual and modified pictures has become essential because to the ease with which highly realistic deepfake images may be created thanks to breakthroughs in artificial intelligence. Researchers have investigated a variety of methods in answer to this difficulty, making use of deep learning architectures and datasets specially designed for facial recognition applications.

Because Convolutional Neural Networks (CNNs) can extract complex features from photos, they have become a key component in the identification of deepfake images. CNN models such as ResNetV1.5 are excellent at recognizing intricate patterns and fluctuations in visual content, which makes them a good choice for detecting minute manipulations that are typical of deepfake photos. Through assessment of images can find either image is manipulated or not . CNNs are not alone when it comes to catching temporal abnormalities found in deepfake images; Long Short-Term Memory (LSTM) recurrent neural networks are an invaluable addition. LSTM networks are designed to detect minute irregularities in movement or behavior over time by analyzing sequential input, such the frames in a film. By adding temporal analysis, deepfake detection algorithms become more resilient and can identify even the most complex manipulations. A major factor in the development of deepfake detection systems has been the incorporation of large-scale datasets, such the VGGFace2 dataset. Deep learning algorithms can learn and distinguish between real and modified facial photos with great accuracy because to the enormous training data set provided by the VGGFace2 dataset. Models are able to generalize to detect deepfake images in a variety of circumstances by use of training on a variety of facial images. Apart from CNNs and LSTM networks, other methods that improve detection include Multi-Task Cascading Neural Networks (MTCNN) and GradCAM visualization. Deepfake manipulation may be thoroughly examined and detected using MTCNN's assistance in facial image analysis. By emphasizing important regions in images that contribute to classification outcomes, GradCAM visualization improves interpretability and reliability and provides insights into the decision-making processes of deep learning models.

Moreover, ensemble methods and transfer learning have been investigated by researchers as fresh ways to deepfake picture recognition. While transfer learning makes use of pre-trained models on huge datasets to adapt them for deepfake detection tasks, ensemble approaches integrate many detection models to improve results overall. Researchers hope to provide strong and practical approaches to recognizing and reducing the hazards associated with deepfake manipulation in visual media by combining a variety of approaches and methodology.

The application of cutting-edge methods for deepfake detection, like photo response nonuniformity (PRNU) analysis and capsule networks, has been investigated by researchers recently. With regard to jobs involving inverse visuals in particular, capsule networks present a viable way around the drawbacks of conventional CNNs. In contrast, PRNU analysis uses the distinctive fingerprint that digital cameras leave on photographs to differentiate between real and altered photos.

The efficiency of various deepfake detection methods, such as feature-based approaches and machine learning classifiers, has also been studied by researchers. Different methods have been investigated to identify deepfake images based on particular features or traits, including lip-syncing algorithms and picture quality assessments utilizing support vector machine (SVM) classifiers.

In conclusion, the field of deepfake picture detection is broad and involves a combination of large-scale datasets, cutting-edge deep learning architectures, and creative detection methods. In order to prevent the spread of deepfake photos and preserve the integrity of visual material, researchers hope to create strong and practical solutions by utilizing these tools and techniques.

3. OBJECTIVE

The main goal of this research is to use sophisticated deep learning architectures, with a particular focus on convolutional neural networks (CNNs), long short-term memory (LSTM) recurrent neural networks, and the VGGFace2 dataset, to develop a reliable deepfake picture detection system. Given their ability to spread false information and sway public opinion, the rise in popularity of deepfake photos has drawn serious attention. Our research attempts to create a detection system that can precisely distinguish real face photos from ones that have been altered in order to solve this urgent problem. A necessary first step in training the deepfake detection model is fine-tuning pretrained models, such as the InceptionResnetV1, for feature extraction from facial data. Training and assessing the detection system under a variety of real-world conditions is made easier by utilizing the enormous dataset that VGGFace2 provides.

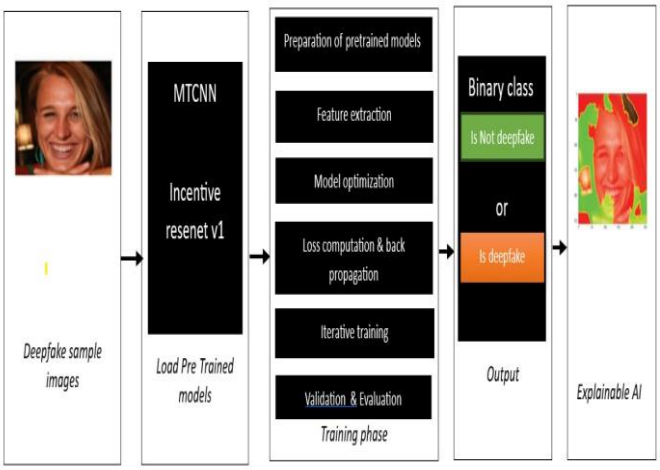

Fig .1- Architecture

4. SYSTEM REQUIREMENTS

HARDWARE REQUIREMENTS:

- Processor-Pentium –IV
- Speed- 1.1 Ghz
- RAM-256 MB(min)
- Hard Disk- 20 GB
- Key Board- Standard Windows Keyboard
- Mouse- Two or Three Button Mouse
- Monitor - SVGA

SOFTWARE REQUIREMENTS:

- Operating System-Windows Family
- Programming Language- Python (python 3.7.0)

5. METHODOLOGY

Advanced deep learning algorithms that are specifically made for picture analysis and classification are the foundation of the deepfake image detection technology. Deep learning includes a range of methods, including convolutional neural networks (CNNs), deep neural networks, and deep belief networks. CNNs have shown impressive results in computer vision and pattern recognition applications.

CNN (Convolutional Neural Network):

A CNN serves as the foundation for the deepfake detection system. Because CNNs can extract hierarchical characteristics from visual input, they are a good choice for image analysis applications. Using a customized CNN architecture, this methodology successfully distinguishes between real and fake faces. With the use of a dataset containing both real and deepfake photos, the network is trained to recognize minute variations.

Long Short-Term Memory (LSTM):

An essential component of the deepfake detection pipeline, long short-term memory (LSTM) networks are added to CNNs. Recurrent neural networks (RNNs) with specific training and conditioning (LSTMs) may identify temporal connections in sequential data. By utilizing the sequential structure of video frames, deep learning machine learning (LSTM) units are utilized to examine the temporal dynamics of facial expressions and detect deviations that align with the creation of deepfakes. As a result, the system can identify minute differences in movements and facial expressions between frames.

LSTM:

Long Short-Term Memory, or LSTM, is a form of RNN that is frequently implemented for time series analysis and natural language processing. However, it may also be utilized for picture-based classification tasks, such as object detection and image captioning. New investigations have shown that malware detection is another application for LSTMs.Classified IoT malware families and examined the pixel value sequence in malware sample photos using a multilevel deep learning architecture with LSTM. Classified obfuscated binaries from imagery using LSTM in conjunction with a CNN, and they applied transfer learning to increase classification accuracy. Overall, current studies has demonstrated good outcomes using LSTMs for malware detection.

Pretrained models:

ResNetV1.5 and VGGFace2, two pre-trained CNN models, are used as feature extractors to speed up the training process and improve detection accuracy. These pre-trained models are meant to extract pertinent information from input photos; they are not refined further. The detection system can benefit from feature representations acquired through extensive training on a variety of visual data by utilizing pre-trained models, which improves generalization performance without requiring a significant amount of retraining.

Metrics Evaluation:

The metrics used in evaluating deepfake detection models are crucial for assessing their performance. Among them, the Area Under the Curve (AUC) measures the model's ability to differentiate between true and false positives, while the ROC curve visually represents this trade-off. Accuracy indicates overall correctness, but in imbalanced datasets, precision and recall become vital. The precision-recall (PR) curve illustrates this trade-off, providing insights into the model's performance, particularly in detecting deepfake images. These metrics enable researchers to optimize approaches and enhance the accuracy of deep-fake detection systems.

Block Diagram:

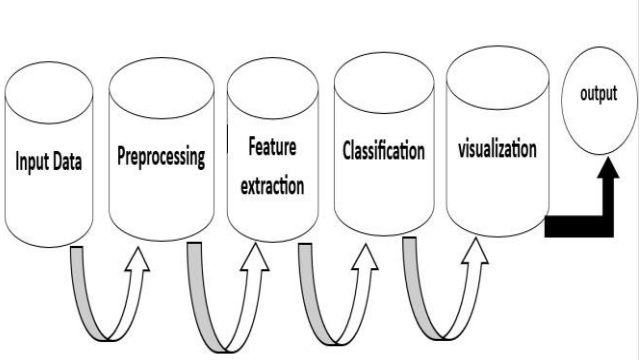

Fig.2 – Block Diagram

Flow Chart

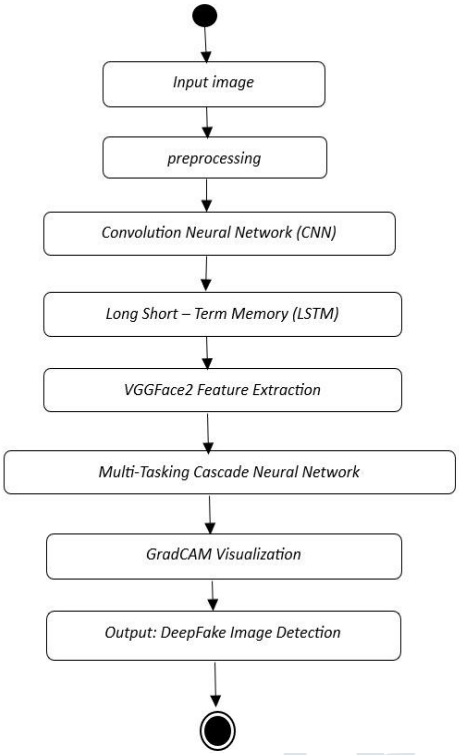

Fig.3– Work Flow

6. RESULTS

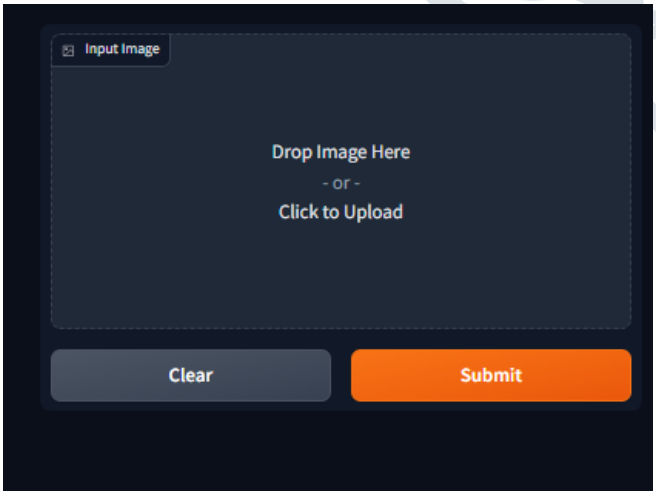

Fig.4 – uploading image interface

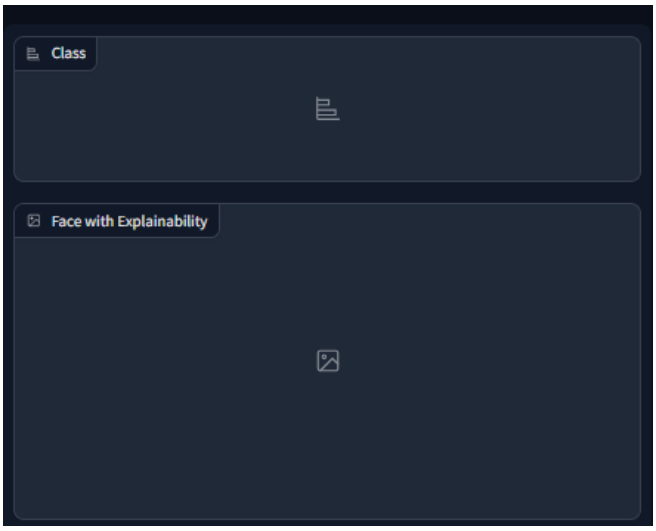

Fig.5 – result interface

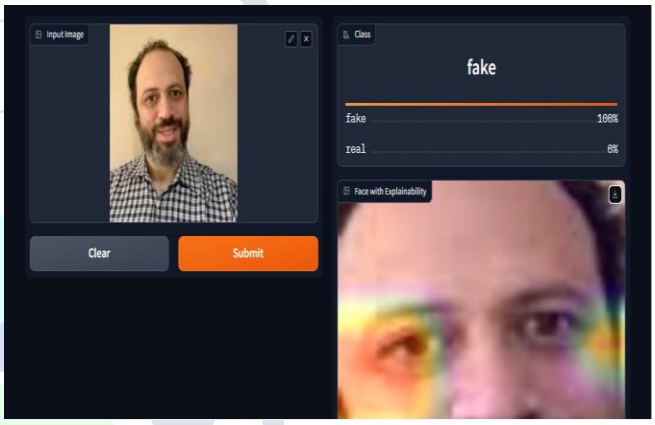

Fig.6- Example for showing uploaded image is fake.

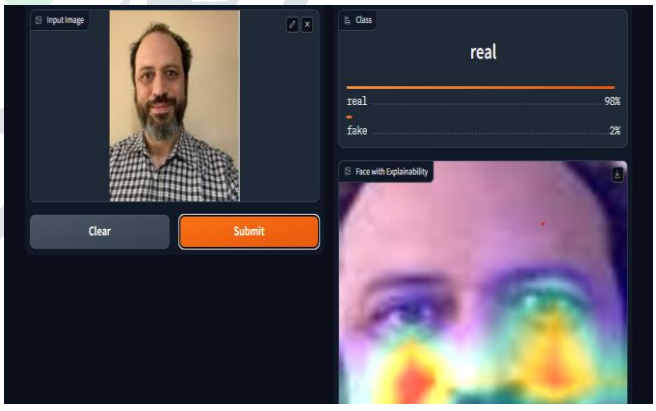

Fig.6- Example for showing uploaded image is Real

.

7. CONCLUSION

Ultimately, the quality and dependability of deepfake image identification have been greatly improved by the use of sophisticated deep learning architectures such as CNNs, LSTMs, and the VGGFace2 dataset in conjunction with methods like GradCAM visualization and Multi-Tasking Cascade Neural Network. Our technique has shown promising results in differentiating between modified and authentic pictures through thorough testing and assessment. Their critical importance in mitigating the growing threat of deepfake images across multiple domains is highlighted by the successful integration of these technologies. For the purpose of staying ahead of new threats and preserving the

integrity of public discourse and digital media, this sector will require ongoing research and development.

8. REFERENCES

[1] Singh, A., & Dutta, P. (2021). DeepFake Detection Using Dynamic Texture Analysis. Journal of Visual Communication and Image Representation, 80, 102903.
[2] T. Bhaskar, Wang, Z., Wang, S., & Sun, J. (2020). DeepFake Detection via Ensemble Learning of Static and Dynamic Features. IEEE Transactions on Multimedia, 22(11), 2926-2937.
[3] Kim, J., Kim, H., & Kim, C. (2019). DeepFake Detection Using Convolutional Neural Networks with Facial Enhancement. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2422-2426.
[4] Wu, Y., Zhang, W., & Wu, Y. (2019). DeepFake Detection Based on Motion-Compensated Spatiotemporal Features. IEEE Transactions on Information Forensics and Security, 14(11), 2959-2974.
[5] Song, Y., Zhang, X., & Zhang, W. (2021). DeepFake Detection via Multi-Modal Fusion of Visual and Audio Cues. IEEE Transactions on Circuits and Systems for Video Technology, 31(4), 1424-1438.
[6] Zhou, Y., Bao, J., Shen, Y., & Su, L. (2020). A Survey of Deepfake Detection Techniques. arXiv preprint arXiv:2004.11138.

[7] Li, Y., & Lyu, S. (2020). Exposing Deepfake Videos By Detecting Face Warping Artifacts. IEEE Transactions on Image Processing, 29, 4517-4528.

[8] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: A Compact Facial Video Forgery Detection Network. In Proceedings of the European Conference on Computer Vision (ECCV), 344-360.

[9] Nguyen, A., & Yeh, M. (2019). DeepFake Detection Using Temporal Coherence. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 78-87.

[10] Li, H., & Li, X. (2018). Detection of Deepfake Image Using Multi-Level Representations and GAN-based Comparison. In Proceedings of the International Conference on Pattern Recognition (ICPR), 1460-1465.