IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Instinctive Authentication Of Ai Generated Content Using LSTM-CNN And Restnet Models

[1]Prof. Manzoor Ahmed, [2]Prof. Ramya H

[1, 2]Assistant Professor

[1, 2] Department of Artificial Intelligence and Machine Learning,

[1, 2] Sri Krishna Institute of Technology, Bangalore, India

Abstract: Our research introduces a robust method to address the growing prevalence of deepfake content, which poses a significant threat to the integrity of multimedia. We present an integrated approach that encompasses both the generation and detection of deepfake technology. Our system leverages state-of-the-art few-shot learning techniques to create personalized and highly realistic talking head models from a limited number of photographs. By employing deep Convolutional Neural Networks (ConvNets) trained on a vast video dataset, our system can produce convincing video sequences mimicking facial expressions and vocal nuances from just a single image. Through extensive meta-learning and adversarial training, our system initializes the parameters of both the generator and discriminator in a person-specific manner, facilitating rapid adaptation and training despite the intricacies of the task. Building upon this foundation, we propose a novel deepfake detection framework that integrates convolutional neural networks (CNNs) to capture temporal dependencies, residual networks (ResNets) for extracting spatial features, and long short-term memory (LSTM) networks. This hybrid architecture effectively combines LSTM-CNN's ability to recognize dynamic facial expressions and movements across frames with ResNet's proficiency in capturing complex facial patterns and contextual information. Furthermore, transfer learning techniques, including pre-training on a diverse dataset and fine-tuning on deepfake-specific data, are utilized to enhance model generalization.

Keywords: Residual Networks (ResNet), Convolutional Neural Networks (ConvNets), and Deepfake Detection.

I. INTRODUCTION

While deep learning holds transformative potential, it has also sparked controversy through applications like deepfake technology. Deep learning, a subset of machine learning, has gained significant attention for its capacity to extract insights from complex data using artificial neural networks. These layered structures enable learning from diverse and unstructured datasets, driving advancements across various domains. However, deepfake technology, which involves the automated creation of synthetic media often altering video content, has raised concerns due to its potential for misuse such as spreading misinformation and cyberbullying. To tackle these challenges, our project proposes an integrated system. Central to our approach is a face forensics model combining innovative detection methods. By leveraging expertise in deep learning and picture forensics, we aim to mitigate the harmful effects of synthetic media. Our strategy blends conventional image forensic techniques with state-of-the-art approaches to identify manipulated facial photos. Additionally, we introduce a convolutional method to detect signs of media manipulation, utilizing the strengths of convolutional neural networks (CNNs) in image analysis. Through comprehensive assessment and validation, our solution aims to locate and prevent the dissemination of deepfake content effectively, addressing associated risks with reliability and efficacy.

1.1 DEEPFAKE CREATION

Traditionally, convolutional neural networks (CNNs) have relied on large datasets of photographs of individuals to generate highly realistic human head images. However, our proposed method aims to overcome this limitation by employing few-shot learning strategies. Our system is designed to learn from minimal data inputs and swiftly adapt to new individuals through meta-learning on a vast dataset of videos. This approach enables the creation of personalized talking head models using just a single image or a small number of image views. Our method initializes the parameters of the generator and discriminator in a person-specific manner, incorporating adversarial training procedures. Even with tens of millions of parameters to adjust, this customized initialization enhances training efficiency with sparse image data. Consequently, our method facilitates rapid convergence and adaptability, ensuring the generation of incredibly lifelike talking head models for new individuals. Experimental results demonstrate the effectiveness of our approach in creating personalized talking head models that perform comparably to more traditional techniques. Moreover, our method offers notable advantages in terms of efficiency and scalability, making it a viable choice for various applications such as entertainment, virtual assistants, and human-computer interaction.

1.2 DEEPFAKE DETECTION

Deepfake and Face2Face exemplify AI algorithms that generate increasingly realistic fake facial content, thanks to advancements in computer vision and deep learning. These technologies manipulate facial expressions or identities, creating media nearly indistinguishable from authentic content. While initially intended for amusement, their misuse has led to significant issues and societal discontent. To address these challenges, we propose a hybrid face forensics framework leveraging convolutional neural networks (CNNs), a type of deep learning model. Our approach enhances manipulation detection by combining two distinct forensic techniques within a single CNN architecture. By drawing insights from both cutting-edge and traditional image forensic methods, we provide a robust tool to combat the spread of falsified media and mitigate associated societal risks. Our hybrid framework excels in detecting subtle signs of manipulation, even within highly realistic fake facial material, making it suitable for real-world applications where prompt and accurate detection is essential. Through comprehensive assessment and testing across various scenarios, we demonstrate the effectiveness and versatility of our method in identifying altered content. In summary, our hybrid face forensics framework represents a significant advancement in manipulation detection, leveraging the strengths of diverse forensic methodologies.

2. LITERATURE SURVEY

[1] In a study, Jee-Young Sun et al. introduced a novel CNN-based approach for contrast enhancement (CE) forensics, demonstrating superior performance compared to traditional methods in detecting forgeries. By integrating gray-Level Co-Occurrence Matrix (GLCM) features, their method exhibits improved accuracy, especially against counter-forensic attacks. This underscores the efficacy of CNNs in CE forensics, offering advanced forgery detection capabilities.

[2]Andreas Rössler et al. introduced "FaceForensics: A comprehensive video dataset for detecting facial forgeries" by harnessing deep learning and Face2Face technology. Their dataset, comprising over 500,000 frames from 1004 videos, encompasses various manipulated scenarios including Source-to-target and Self-reenactment manipulations. This dataset significantly surpasses existing collections and serves as a valuable resource for research in forgery detection.

[3] "MetaGAN: An innovative approach to Few-Shot Learning" by Ruixiang Zhang et al. introduces MetaGAN, a novel framework designed to address few-shot learning challenges. MetaGAN offers a straightforward yet powerful method to enhance the performance of few-shot learning models, demonstrating versatility and flexibility in tackling such tasks.

[4]A. Bromme et al. conducted a study titled "Evaluation of Fake Face Detection Methods: Assessing Generalization Abilities". They evaluated various methods including Local Binary Patterns (LBP) and CNN models such as AlexNet and ResNet50 for fake face detection. Despite not being explicitly trained for this

task, CNN models exhibited superior performance compared to other methods, showcasing their potential for identifying fake faces despite evolving technology.

3. PROBLEM STATEMENT

The rapid advancement of deepfake technologies poses a pervasive threat to the authenticity of multimedia content, sparking concerns regarding disinformation and fraudulent activities. Existing detection systems struggle to keep pace with the evolving sophistication of deepfake techniques, leaving digital media vulnerable to manipulation and misuse. The complexity of temporal dynamics in video sequences and the intricate spatial nuances of facial features further compound the challenge. Consequently, there is an urgent need for novel and robust methods to detect deepfake content promptly and accurately. It is imperative to develop practical solutions to combat the proliferation of deepfakes, as failure to do so could undermine the integrity of digital media and diminish public trust in online information sources.

3.1 EXISTING SYSTEM

One method for detecting deepfakes involves scrutinizing visible anomalies in physical or physiological attributes within images or videos. Researchers analyze factors such as irregular shadows, distorted geometry, or discrepancies in facial features like teeth, ears, and eye colors to ascertain authenticity. For instance, Li et al. utilized eye blink patterns in videos to flag potential abnormalities indicative of deepfake manipulation. Similarly, other techniques concentrate on disparities between head and body movements to uncover patterns challenging for deepfake algorithms to replicate accurately. These efforts seek to pinpoint irregularities that distinguish genuine content from synthetic counterparts.

3.2 PROPOSED SYSTEM

Our proposed approach entails integrating convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, alongside residual networks (ResNet), to model temporal relationships and extract spatial features. This hybrid architecture aims to leverage the complementary strengths of LSTM-CNN, adept at capturing dynamic temporal changes in video sequences, and ResNet, renowned for its ability to discern intricate spatial patterns. We plan to employ transfer learning techniques, initially pre-training the model on diverse datasets to establish a robust foundation. Subsequently, fine-tuning will be conducted on datasets specific to deepfake generation, enhancing adaptability to evolving deepfake techniques. By doing so, our approach endeavors to address the limitations of current methods, offering a comprehensive and adaptable solution to the intricate challenges of deepfake detection.

4. IMPLEMENTATION

4.1 METHODOLOGY FOR CREATION OF FAKE IMAGE

Although Photoshop and After Effects are commonly used by professionals every day, merely installing these programs does not guarantee the ability to create photorealistic images and videos. Similarly, producing realistic face-swapped videos presents significant challenges. The initial iteration of deepfake technology, pioneered by a Reddit user through FakeApp, employed an autoencoder-decoder fusion architecture. This method requires two sets of encoder-decoder pairs to exchange faces between source and target images, with each pair trained on a corresponding image set while sharing encoder parameters. Notably, FakeApp utilizes Google's AI Framework TensorFlow, which was previously employed in projects like DeepDream. Alternative open-source options to FakeApp include DeepFaceLab, FaceSwap, and FakeApp. Regardless of the application chosen for generating deepfakes, The process typically involves three main steps: extraction, training, and creation.

Extraction:
The origin of the term "deep" in deepfake stems from its reliance on extensive datasets within deep learning. Generating a deepfake video entails the production of thousands of distinct images. This production process involves extracting all frames, identifying faces, and aligning them, termed as the

extraction procedure. A pivotal aspect of this procedure is ensuring alignment, wherein all faces are standardized to the same size as the neural network undergoes replacement.

Training:

The term "training" originates from Machine Learning and denotes the process enabling a neural network to transform one face into another. Despite its time-consuming nature, the training phase needs to be conducted only once. Upon completion, it empowers the conversion of a face from individual A to individual B.

Creation:

Upon completion of training, the next step is the creation of a deepfake. Whether starting from a video or an image, all frames are extracted and facial alignments are ensured. Subsequently, each face is transformed using the trained neural network. The final step involves reintegrating the transformed face back into the initial frame. Despite its apparent simplicity, this stage is where many face-swapping applications encounter challenges. As mentioned before, autoencoders play a crucial role in deepfake creation.

ARCHITECTURE OF DEEPFAKE CREATIONMODEL

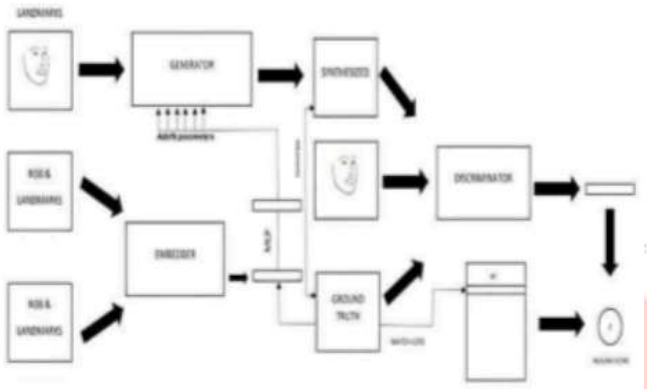

Fig 4.1 A proposed creation model

Once the training concludes, the next step involves creating the deepfake. Whether starting with a video or an image, all frames are extracted and facial features are aligned. Subsequently, each individual is transformed using the trained neural network.

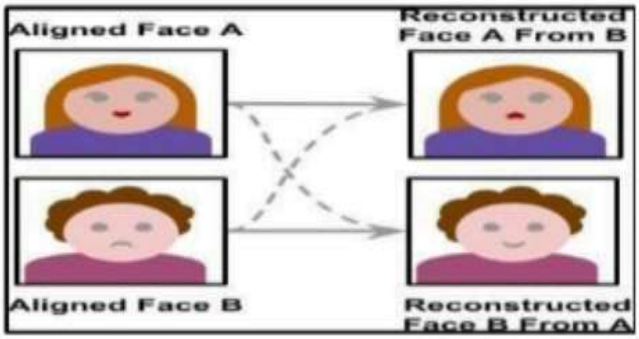

Fig 4.2 Training of an image from face A face B

4.2 METHODOLOGY FOR DETECTION OF FAKE IMAGE

Deepfakes represent a form of synthetic media that involves replacing an individual's face within existing images or videos with the likeness of another person. This process typically relies on an encoder-decoder model, where the encoder gathers data from both the desired target face and the original source face images. To accommodate the usual brevity of deepfakes, the video is often divided into segments, allowing the decoder unit to utilize these features to generate a fabricated video featuring the desired face. Advanced processing techniques are then applied to refine the video quality and eliminate visible artifacts, although subtle remnants may remain imperceptible to the naked eye. Interestingly, these subtle remnants serve as crucial cues for our detection model, which employs the 5 recursive neural network inceptionresnetv2 for feature extraction. While deepfakes leverage cutting-edge facial manipulation methodologies such as generative adversarial networks (GANs) and autoencoders, our detection model focuses on identifying these

minute traces left behind during the manipulation process. This field presents a captivating blend of creative possibilities and ethical considerations.

ARCHITECTURE FOR DETECTION OF DEEPFAKE MODEL

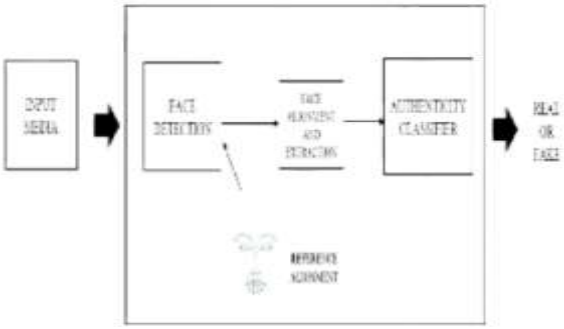

Fig 4.3 A proposed detection model

Dataset and Preprocessing:

The dataset underwent meticulous curation from diverse sources, including the datasets from the Kaggle Deepfake Detection Challenge, Face Forensics, and Celeb-Deepfake Forensics. In total, it comprised 401 videos, intriguingly featuring both authentic footage and manipulated content. The manipulation process involved paid actors modifying real videos, subsequently transformed into deepfakes through various generator methods. For our model, we partitioned the dataset into a 70% training set and a 30% testing set. During the training phase, we provided the machine with labels corresponding to the video files, crucially identifying the specific frame where the original video transitioned into a deepfake. This frame underwent thorough analysis during preprocessing. Generally, due to computational constraints, we could only extract 147 frames from each video during preprocessing. These frames were further subdivided into smaller batches for training and testing. Our objective was to develop an efficient deepfake detection system capable of discerning subtle manipulations within videos.

Modelling Model

The system performs image categorization analysis on every frame extracted from the video. We utilized a pretrained CNN model called Inception ResNetV2 [12], alongside RNN and LSTM. Additionally, we defined the loss function, optimizer, and other hyperparameters necessary for the training process. Depending on the training model's state, the learning rate should be dynamically adjusted to minimize the loss value.

i) Face detection

An input image undergoes facial zone detection through a neural facial landmark detection model, which autonomously identifies 68 fiducial facial landmark points surrounding facial features and contours like eyes, mouth, and chin. Of these points, only 51 are utilized, excluding the 17 points from the chin, as facial manipulation focuses on the inner facial area.

ii) Face alignment and extraction

Subsequently, the system adjusts the face to conform to the reference alignment since faces in media often deviate from frontal or unrotated positions. Employing an affine transformation on the image, we establish a one-to-one mapping from the extracted landmark points to the reference alignment points. This transformation effectively aligns rotated or profile faces with the reference alignment, thereby improving the performance of fake face detection. Finally, the system extracts the facial region from the image and inputs it into the facial authenticity classifier.

iii) Authenticity classification

The proposed face authenticity classifier integrates a 3 content feature extractor (CFE) and a trace feature extractor (TFE). Each convolutional operation is represented by a square that elaborates its details within the two feature extractors. For instance, the initial convolution in the CFE utilizes a 7 × 7 convolutional filter with a stride of 2, yielding 64 feature maps as output.

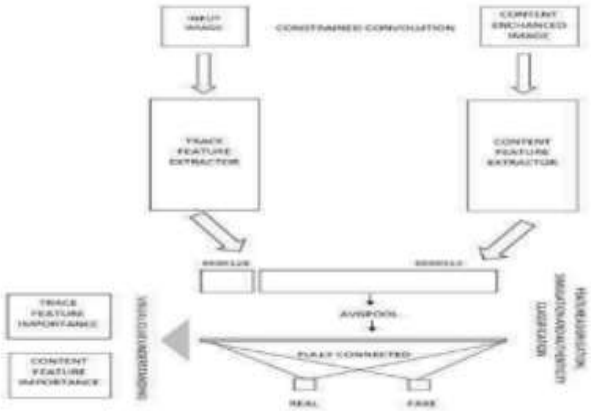

Fig 4.4 Authenticity Classifier

5. CONCLUSION

Advancements in artificial intelligence, particularly in deep learning, have led to the emergence of deepfake technology, enabling the creation of highly realistic multimedia content aimed at deceiving viewers. This poses numerous risks, including dissemination of false information, harm to individual reputations, and even threats to national security. Given the growing threat of deepfakes, the importance of developing effective detection techniques is paramount. The novel deepfake detection method proposed in this paper integrates Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and Residual Networks (ResNet), effectively addressing the limitations of prior approaches by capturing both spatial and temporal data. Transfer learning is employed to enhance generalization and adaptability of the model to the rapidly evolving deepfake landscape.

The proliferation of deepfakes has exacerbated the challenge of distinguishing between authentic and manipulated content, eroding public trust in media. They pose significant risks to political stability, fueling hate speech and misinformation, particularly on social media platforms where they can rapidly propagate and target specific demographics. Consequently, media literacy and the ability to identify manipulated content become crucial, especially in legal contexts where digital evidence is presented. The suggested hybrid forensic framework, utilizing a convolutional neural network, integrates facespecific analysis with general-purpose image forensics to provide robust defense against deepfake threats. Experimental results validate the efficacy of this approach, underscoring the need for ongoing research in deepfake detection to safeguard privacy and uphold global, political, and societal security. Future research on the proposed model aims to reduce computational overhead and enhance speed by exploring methods such as processing high-resolution images without downsampling using neural networks equipped with global pooling. Additionally, investigating various image upscaling techniques and assessing their impact on performance could offer valuable insights for improving accuracy while managing trade-offs with time latency.

6. REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.