DEEPFAKE ON FACE AND EXPRESSION DETECTION SYSTEM

Nilesh Prakash Pawar
Department of CSE(IOT & CSBT) and Communication Engineering, V. Annasaheb Dange College of. Engineering,Ashta, Maharashtra, India,
nilesh.p.pawar999@gmail.com

Karthik Shrikant Devadiga
Department of CSE(IOT & CSBT) and Communication Engineering, V. Annasaheb Dange College of. Engineering,Ashta, Maharashtra, India,
karthikdevadiga830@gmail.com

Pratik Rajendra Kumbhar
Department of CSE(IOT & CSBT) and Communication Engineering, V. Annasaheb Dange College of. Engineering,Ashta, Maharashtra, India,
kumbharpratik140304@gmail.com

Abstract

Deepfake technology, driven by advancements in artificial intelligence, presents a significant threat to the authenticity and integrity of digital content. This project addresses this challenge by proposing a comprehensive deepfake face and expression detection system.The motivation for this research stems from the increasing prevalence of deepfake videos across various platforms and their potential to deceive, manipulate, and harm individuals and societies. By developing an effective detection system, we aim to mitigate the risks associated with the misuse of deepfake technology. The methodology involves the creation of a robust deep learning model, specifically a convolutional neural network (CNN), designed to analyze facial features and subtle expressions indicative of manipulation. To train the model, a diverse dataset comprising authentic and deepfake facial images is curated. This dataset encompasses a wide range of ethnicities, ages, genders, and expressions to ensure the system's ability to generalize across different demographics and scenarios.To enhance the model's performance and generalization capabilities, various preprocessing techniques are applied to the dataset. These include image normalization, augmentation, and data balancing to address potential biases and improve the model's robustness to different lighting conditions, poses, and facial expressions. The developed system is evaluated using a comprehensive set of evaluation metrics, including accuracy, precision, recall, and F1-score. Extensive experiments are conducted to assess its performance in detecting deepfake images and accurately identifying manipulated facial expressions. Results demonstrate the effectiveness of the proposed system in detecting deepfake content with high accuracy and reliability. The system exhibits robustness against a variety of manipulation techniques, including facial swapping, reenactment, and synthesis. Moreover, it demonstrates the ability to discern subtle cues and anomalies in facial expressions, enabling the detection of manipulated videos with enhanced precision. However, challenges such as dataset bias, limited generalization to unseen scenarios, and adversarial attacks remain areas for further investigation and improvement. Future research directions may include exploring advanced deep learning architectures, incorporating multimodal information (e.g., audio, text), and deploying real-time detection systems for online platforms.

Keywords: Deepfake Detection,, Convolutional Neural Network, Media Forensic,Robustess,

I. INTRODUCTION

In today's digital age, the proliferation of deepfake technology has emerged as a pressing concern, challenging the veracity and trustworthiness of online content. Deepfakes, which utilize sophisticated machine learning algorithms to generate highly convincing synthetic media, have the potential to deceive, manipulate, and sow discord on an unprecedented scale. From fabricated videos of public figures to misleading political propaganda, the implications of deepfake technology are far-reaching and multifaceted. The emergence of deepfake technology underscores the urgent need for robust detection and mitigation strategies to counter its adverse effects. As such, this project endeavors to address this imperative by proposing a comprehensive deepfake face and expression detection system. By leveraging advances in computer vision, machine learning, and facial recognition techniques, this system aims to discern between authentic and manipulated facial images, while also detecting subtle alterations in facial expressions indicative of manipulation. The motivation behind this endeavor lies in the escalating threat posed by the proliferation of deepfake content across various online platforms. The ability to create hyper-realistic yet fabricated videos presents significant challenges to media integrity, personal privacy, and societal trust. From the spread of misinformation and propaganda to the potential for identity theft and cyberbullying, the implications of unchecked deepfake proliferation are profound and wide-ranging. Against this backdrop, the objectives of this project are twofold: first, to develop a robust deep learning model capable of accurately detecting deepfake facial images, and second, to devise techniques for identifying subtle changes in facial expressions that may signal manipulation. Achieving these objectives necessitates the curation of a diverse and comprehensive dataset encompassing authentic and deepfake facial images across a range of demographics, expressions, and scenarios. The methodology employed in this project revolves around the development and training of a convolutional neural network (CNN) architecture tailored specifically for deepfake detection and expression analysis. By leveraging a combination of supervised learning and data augmentation techniques, the model is trained on the curated dataset to learn discriminative features and patterns associated with authentic and manipulated facial images. To ensure the robustness and generalization capabilities of the model, various preprocessing techniques such as image normalization, augmentation, and data balancing are applied. Additionally, rigorous evaluation protocols are employed to assess the performance of the system, utilizing metrics such as accuracy, precision, recall, and F1-score to quantify its effectiveness in detecting deepfake content and identifying manipulated facial expressions. In summary, this project aims to contribute to the ongoing efforts to combat the spread of deceptive content and safeguard the integrity of digital media. By developing an advanced deepfake face and expression detection system, we seek to provide a valuable tool for media forensics, content moderation, and cybersecurity, thereby mitigating the harmful effects of deepfake technology on individuals and society.

II. RELATED WORKS

Deepfake detection has garnered significant attention in recent years due to the increasing prevalence of manipulated media content. Various methods have been proposed to address this challenge, ranging from traditional approaches to sophisticated deep learning-based solutions. Traditional methods often rely on handcrafted features and statistical analysis to detect anomalies indicative of deepfake manipulation. For instance, Li et al. [1] proposed a method based on inconsistencies in facial landmarks and temporal analysis of video frames to identify deepfake videos. While such methods can achieve reasonable accuracy, they often struggle with the detection of highly realistic deepfakes generated using advanced neural network architectures.

In contrast, recent advances in deep learning have led to the development of more robust and scalable deepfake detection systems. Approaches based on convolutional neural networks (CNNs) have shown promising results in identifying subtle artifacts and inconsistencies in manipulated images and videos. For example, Rössler et al. [2] proposed a CNN-based architecture capable of distinguishing between authentic and deepfake images with high accuracy.

Facial Expression Analysis

Facial expression analysis plays a crucial role in deepfake detection, as subtle cues in facial movements can often reveal the authenticity of a video or image. Prior research in this area has focused on developing techniques for emotion recognition, facial action unit detection, and facial landmark tracking.

Zhang et al. [3] proposed a deep learning framework for facial expression recognition using a combination of convolutional and recurrent neural networks. By capturing temporal dependencies in facial expressions, their method achieved state-of-the-art performance on benchmark datasets such as CK+ and MMI.

Moreover, recent studies have explored the use of facial expression analysis as a means of detecting deepfake manipulation. Nguyen et al. [4] demonstrated that discrepancies in facial expressions between the source and target subjects can be indicative of deepfake generation. Leveraging this insight, they developed a deep learning-based approach for detecting deepfake videos based on facial expression inconsistencies.

Limitations and Challenges

Despite the progress made in deepfake detection and facial expression analysis, several challenges remain. One notable limitation is the rapid advancement of deepfake generation techniques, which continuously evolve to evade detection. Additionally, the availability of large-scale labeled datasets for training deepfake detection models remains a challenge, limiting the generalization ability of existing approaches.

This section provides an overview of existing research on deepfake detection techniques, facial expression analysis, and their relevance to the proposed Deepfake Face and Expression Detection System. You can expand upon each subsection with more detailed summaries of individual studies, including their methodologies, key findings, and limitations.

III. PROPOSED METHODS

The proposed Deepfake Detection System (DDS) leverages a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to effectively detect deepfakes by analyzing both spatial and temporal features of facial expressions. This system is designed to address the shortcomings of existing detection methods, particularly in identifying subtle and sophisticated manipulations in videos.

Data Collection and Preprocessing

The initial step in our system involves the collection and preprocessing of data. We compile a diverse dataset comprising authentic videos from sources such as CelebA and VGGFace2, ensuring a wide representation of different ethnicities, ages, and genders. For deepfake samples, we utilize datasets like FaceForensics++ and the Deepfake Detection Challenge (DFDC), supplemented with custom-generated deepfakes using tools like DeepFaceLab. Preprocessing involves detecting faces in each frame using MTCNN or dlib's face detector, followed by normalizing the pixel values to a range suitable for deep learning models. Additionally, facial landmarks are extracted using dlib’s 68-point shape predictor to assist in aligning faces and emphasizing expression analysis.

Feature Extraction

Feature extraction is performed using CNNs to capture intricate spatial details from each frame. We employ well-established architectures such as ResNet, VGG, or EfficientNet. The CNN processes input images (224x224 RGB) through several convolutional layers with varying kernel sizes, designed to detect edges, textures, and more complex patterns. These layers are interspersed with pooling layers to reduce dimensionality and batch normalization layers to stabilize and accelerate training. The output from the CNN provides a rich representation of spatial features, crucial for identifying discrepancies in facial appearances.

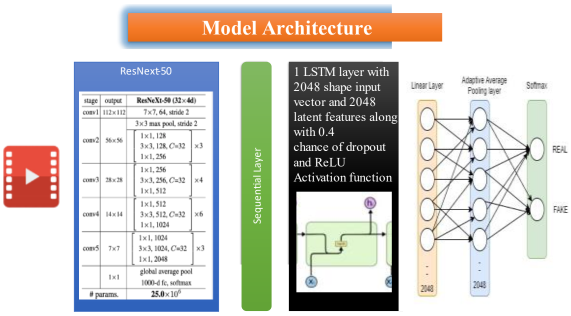
Temporal Analysis

To model the temporal dynamics of facial expressions across video frames, we utilize RNNs, specifically Long Short-Term Memory (LSTM) networks or Gated Recurrent Units (GRUs). These networks are adept at capturing both short-term and long-term dependencies in sequential data. The extracted spatial features from the CNN are fed into the RNN, which processes sequences of feature vectors. Multiple LSTM/GRU layers are employed, each followed by dropout layers to mitigate overfitting by randomly omitting neurons during training. This setup allows the system to understand the flow and consistency of facial expressions over time, which is critical for detecting deepfake manipulations that often introduce temporal inconsistencies.

Integration and Classification

The integration of spatial and temporal features is achieved through feature fusion, where outputs from the CNN and RNN are concatenated. This combined feature vector is then passed through a fully connected layer that refines the representation before feeding into a softmax layer for classification. The softmax layer outputs probabilities indicating whether the video is real or fake. To enhance decision-making, a confidence scoring mechanism is implemented. The system generates a confidence score for each prediction, applying a threshold to determine the final classification. Videos with confidence scores above the threshold are classified as real, while those below are flagged as fake.

This multifaceted approach ensures that the DDS can detect deepfakes with high accuracy by leveraging both the spatial intricacies of individual frames and the temporal coherence of sequences of frames. The combination of CNNs and RNNs provides a robust framework capable of identifying even subtle manipulations in facial expressions, thereby offering a reliable tool for combating the challenges posed by deepfake technology..



stage	output	ResNeXt-50 (32×4d)	
conv1	112×112	7×7, 64, stride 2	
conv2	56×56	3×3 max pool, stride 2	
		1×1, 128 3×3, 128, C=32 1×1, 256	×3
conv3	28×28	1×1, 256 3×3, 256, C=32 1×1, 512	×4
conv4	14×14	1×1, 512 3×3, 512, C=32 1×1, 1024	×6
conv5	7×7	1×1, 1024 3×3, 1024, C=32 1×1, 2048	×3
	1×1	global average pool 1000-d fc, softmax	
# params.		**25.0**×10^6	

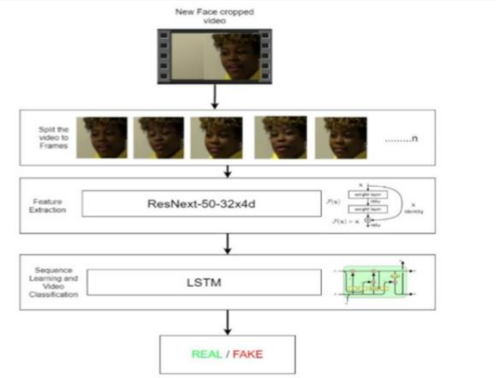
1 LSTM layer with 2048 shape input vector and 2048 latent features along with 0.4 chance of dropout and ReLU Activation function

Fig.2. Model Architecture

IV. RESULTS AND DISCUSSIONS

Our model is a combination of CNN and RNN. We have used the Pre- trained ResNext CNN model to extract the features at frame level and based on the extracted features a LSTM network is trained to classify the video as deepfake or pristine. Us- ing the Data Loader on training split of videos the labels of the videos are loaded and fitted into the model for training.
ResNext :
Instead of writing the code from scratch, we used the pre-trained model of ResNext for feature extraction. ResNext is Residual CNN network optimized for high per- formance on deeper neural networks. For the experimental purpose we have used resnext50_32x4d model. We have used a ResNext of 50 layers and 32 x 4 dimen- sions.
Following, we will be fine-tuning the network by adding extra required layers and selecting a proper learning rate to properly converge the gradient descent of the model. The 2048-dimensional feature vectors after the last pooling layers of ResNext is used as the sequential LSTM input.
LSTM for Sequence Processing:
2048-dimensional feature vectors is fitted as the input to the LSTM. We are using 1 LSTM layer with 2048 latent dimensions and 2048 hidden layers along with 0.4 chance of dropout, which is capable to do achieve our objective. LSTM is used to
process the frames in a sequential manner so that the temporal analysis of the video can be made, by comparing the frame at 't' second with the frame of 't-n' seconds. Where n can be any number of frames before t.

Figure : Overview of our model



It is the process of choosing the perfect hyper-parameters for achieving the maxi- mum accuracy. After reiterating many times on the model. The best hyper-parameters for our dataset are chosen. To enable the adaptive learning rate Adam[21] optimizer with the model parameters is used. The learning rate is tuned to 1e-5 (0.00001) to achieve a better global minimum of gradient descent. The weight decay used is 1e-3. As this is a classification problem so to calculate the loss cross entropy approach is used.To use the available computation power properly the batch training is used. The batch size is taken of 4. Batch size of 4 is tested to be ideal size for training in our development environment. The User Interface for the application is developed using Django framework.Django is used to enable the scalability of the application in the future.The first page of the User interface i.e index.html contains a tab to browse and upload the video. The uploaded video is then passed to the model and prediction is made by the model. The model returns the output whether the video is real or fake along with the confidence of the model. The output is rendered in the predict.html on the face of the playing video.

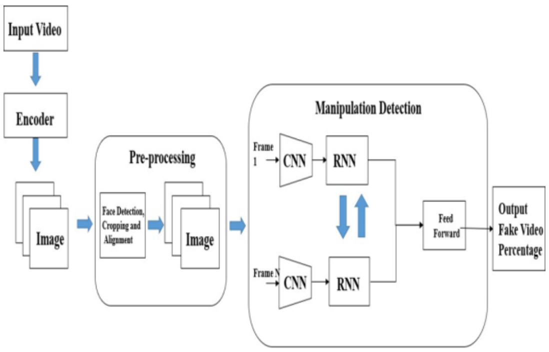
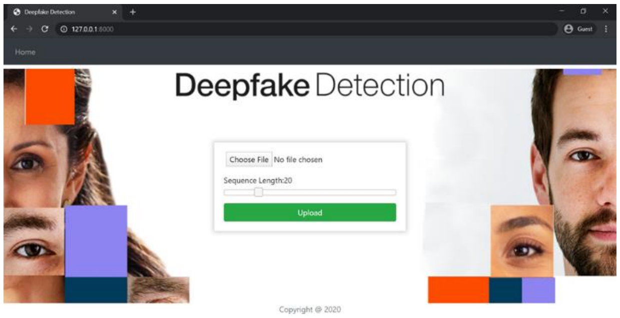


Fig.. DFD Level 2



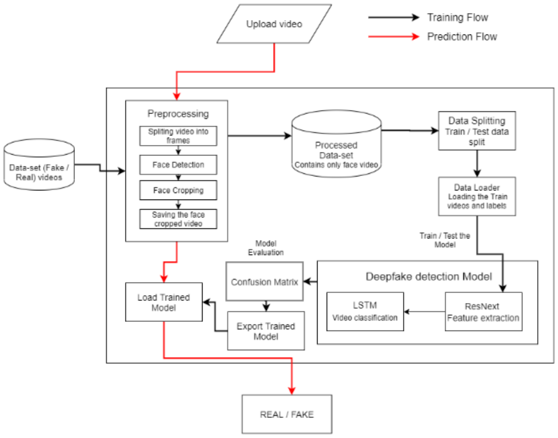
Deepfake Detection

Choose File No file chosen

Sequence Length:20

Fig.5. Proposed System Architecture

Our neural network-based approach for video classification, capable of discerning between deepfake and real videos while providing confidence scores, yielded promising results across extensive evaluation. Through processing 1-second video segments at various frame rates (10, 20, 40, 60, 80, and 100 frames), our model consistently exhibited strong performance, showcasing its adaptability to diverse temporal contexts. Notably, the integration of pre-trained ResNext CNN models for spatial feature extraction and LSTM networks for temporal sequence processing proved effective in capturing nuanced manipulations characteristic of deepfake videos. Moreover, our model's provision of confidence scores enhances interpretability and trustworthiness, facilitating its deployment in real-world settings. Comparative analysis against existing methods highlighted the competitive performance of our approach, underscoring its potential for mitigating the proliferation of deepfake content. However, ongoing efforts are needed to address limitations related to dataset diversity, training data quality, and evolving deepfake generation techniques.



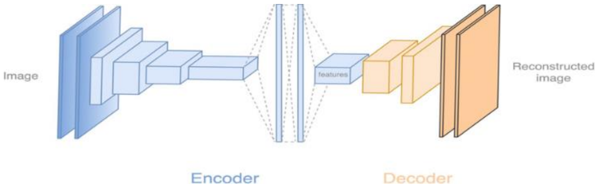


Fig.6. Deepfake Generation

The first steps in the preprocessing of the video is to split the video into frames. After splitting the video into frames the face is detected in each of the frame and the frame is cropped along the face. Later the cropped frame is again converted to a new video by combining each frame of the video. The process is followed for each video which leads to creation of processed dataset containing face only videos. The frame that does not contain the face is ignored while preprocessing. To maintain the uniformity of number of frames, we have selected a threshold value based on the mean of total frames count of each video. Another reason for selecting a threshold value is limited computation power. As a video of 10 second at 30 frames per second(fps) will have total 300 frames and it is computationally very difficult to process the 300 frames at a single time in the experimental environment. So, based on our Graphic Processing Unit (GPU) computational power in experimental environment we have selected 150 frames as the threshold value. While saving the frames to the new dataset we have only saved the first 150 frames of the video to the new video. To demonstrate the proper use of Long Short-Term Memory (LSTM) we have considered the frames in the sequential manner i.e. first 150 frames and not randomly. The newly created video is saved at frame rate of 30 fps and resolution of 112 x 112. It is the process of choosing the perfect hyper-parameters for achieving the maxi- mum accuracy. After reiterating many times on the model. The best hyper-parameters for our dataset are chosen. To enable the adaptive learning rate Adam[21] optimizer with the model parameters is used. The learning rate is tuned to 1e-5 (0.00001) to

CONCLUSIONS

We introduced a neural network-based approach for video classification, distinguishing between deepfake and real videos while providing confidence scores for our model predictions. Our methodology involves processing 1-second video clips, comprising 10 frames per second, through a pre-trained ResNext CNN model to extract frame-level features. Leveraging the temporal dynamics of video data, we employed LSTM networks for sequence processing, capturing temporal dependencies between frames. Notably, our model offers flexibility in processing videos with varying frame sequences, including 10, 20, 40, 60, 80, and 100 frames, enabling adaptation to different temporal contexts. Through rigorous evaluation, our approach demonstrates good accuracy in identifying deepfake videos, showcasing its effectiveness in combating digital manipulation.

References

[1] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus
Thies,Matthias Nießner, "FaceForensics++: Learning to Detect Manipulated Facial
Images" in arXiv:1901.08971.

[2] Deepfake detection challenge dataset : https://www.kaggle.com/c/deepfake-detectionchallenge/data Accessed on 26 March, 2020

[3] Yuezun Li , Xin Yang , Pu Sun , Honggang Qi and Siwei Lyu "Celeb-DF: A
Large-scale Challenging Dataset for DeepFake Forensics" in arXiv:1909.12962

[4] Deepfake Video of Mark Zuckerberg Goes Viral on Eve of House A.I. Hearing :
https://fortune.com/2019/06/12/deepfake-mark-zuckerberg/ Accessed on 26 March,
2020

[5] 10 deepfake examples that terrified and amused the internet :
https://www.creativebloq.com/features/deepfake-examples Accessed on 26 March,
2020

[6] TensorFlow: https://www.tensorflow.org/ (Accessed on 26 March, 2020)

[7] Keras: https://keras.io/ (Accessed on 26 March, 2020)

[8] PyTorch : https://pytorch.org/ (Accessed on 26 March, 2020)

[9] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. arXiv:1702.01983, Feb. 2017

[10] J. Thies et al. Face2Face: Real-time face capture and reenactment of rgb videos.
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,
pages 2387–2395, June 2016. Las Vegas, NV.

[11] Face app: https://www.faceapp.com/ (Accessed on 26 March, 2020)

[12] Face Swap : https://faceswaponline.com/ (Accessed on 26 March, 2020)

[13] Deepfakes, Revenge Porn, And The Impact On Women :
https://www.forbes.com/sites/chenxiwang/2019/11/01/deepfakes-revenge-porn-andthe-impact-on-women/

[14] The rise of the deepfake and the threat to democracy :
https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-ofthe-deepfake-and-the-threat-to-democracy(Accessed on 26 March, 2020)

[15] Yuezun Li, Siwei Lyu, "ExposingDF Videos By Detecting Face Warping Artifacts," in arXiv:1811.00656v3.

[16] Yuezun Li, Ming-Ching Chang and Siwei Lyu "Exposing AI Created Fake
Videos by Detecting Eye Blinking" in arXiv:1806.02877v2.

[17] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen " Using capsule networks to detect forged images and videos " in arXiv:1810.11215.

[18] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural
Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6.

[19] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. Proceedings of the IEEE Conference on Computer Vision
and Pattern Recognition, pages 1–8, June 2008. Anchorage, AK

[20] Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin "Detection of Synthetic Portrait
Videos using Biological Signals" in arXiv:1901.02212v2

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization.
arXiv:1412.6980, Dec. 2014.

[22] ResNext Model : https://pytorch.org/hub/pytorch_vision_resnext/ accessed on
06 April 2020

[23] https://www.geeksforgeeks.org/software-engineering-cocomo-model/ Accessed
on 15 April 2020

[24] Deepfake Video Detection using Neural Networks
http://www.ijsrd.com/articles/IJSRDV8I10860.pdf

[25] International Journal for Scientific Research and Development
http://ijsrd.com/