IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2024

DOI: 10.17148/IJARCCE.2024.131023

Development of deepfakes detection model using deep learning framework

Likith P[1], Suresh.S.R[2], Shraddha. C[3]

Student, Electronics and Communication Engineering, MS Ramaiah University of Applied Sciences, Bangalore, India[1]

Student, Information Science and Engineering, Global Academy of Technology, Bangalore, India[2,3]

Abstract: In response to the growing threat of deepfake technology, which can deceive and manipulate individuals, leading to identity theft, financial fraud, and political manipulation. This paper proposes a deepfakes detection model using a deep learning framework and image processing techniques. The proposed utilizes the ResNeXt architecture as a powerful feature extractor to capture intricate patterns and discriminative features from input images or video frames. Additionally, the LSTM architecture is employed to handle temporal dependencies, enabling the model to analyze the temporal coherence and consistency of video sequences. By leveraging ResNeXt and LSTM, the model achieves enhanced accuracy and robustness in detecting deepfake content.

Keywords: LSTM- long short term memory, cnn- convolution neural network.

I. INTRODUCTION

Deepfake technology, powered by advanced machine learning algorithms, has emerged as a significant technological advancement with both positive and negative implications. Deepfakes are synthetic media, typically videos, that convincingly depict individuals saying or doing things they never actually did. While deepfakes have several legitimate applications, such as entertainment and de-ageing in movies, they also present serious threats to various aspects of society. Deepfakes pose significant privacy violations. By superimposing someone's likeness onto explicit or compromising content, deepfakes can lead to harassment, blackmail, and damage to personal and professional reputations. Privacy breaches of this nature can have severe psychological and emotional consequences for the individuals targeted. Furthermore, deepfakes raise legal and ethical concerns. They can infringe upon intellectual property rights, challenge the authenticity of digital evidence in legal proceedings, and blur the line between reality and fabrication. Ethical dilemmas arise regarding consent, media manipulation, and responsible use of AI, while deepfake technology offers exciting possibilities, its potential for misuse is substantial. Examples of potential misuse include the dissemination of misinformation, fraud and scams, privacy violations, and legal and ethical concerns. It is crucial for individuals, organizations, and policymakers to address these risks through technological advancements, public awareness campaigns, and robust regulatory frameworks to mitigate the negative impact of deepfakes on society. This scientific paper introduces a comprehensive framework, which proposes an innovative approach to address the challenges of deepfake detection. we propose a comprehensive approach that combines a deep learning framework and image processing techniques. The deep learning framework enables us to leverage the power of neural networks for learning complex patterns and features from the input data. Firstly, we gather a diverse dataset consisting of both authentic and deepfake media samples. This dataset is carefully curated to ensure a representative distribution and minimize biases. We preprocess the dataset using image processing techniques, such as resizing, cropping, and normalization, to enhance the quality and consistency of the data. Next, we design and implement a deep learning model architecture suitable for deepfake detection. The ResNeXt architecture is employed as a backbone network for feature extraction. It leverages the power of deep convolutional neural networks (CNNs) to extract high-level features from the input images or video frames. By using ResNeXt, the model can capture intricate patterns and discriminative features that distinguish between authentic and deepfake media. Additionally, the LSTM architecture is utilized to handle the temporal dependencies present in video sequences. LSTM networks are capable of modeling long term dependencies and capturing temporal dynamics, which are crucial for detecting subtle manipulations present in deepfake videos. By incorporating LSTM, the model can effectively analyze the temporal coherence and consistency of the video frames, aiding in the identification of deepfake content. Furthermore, image processing techniques are applied to preprocess the data and enhance the quality of the input media. These techniques involve resizing, cropping, and normalization to standardize the data and improve the clarity and consistency of the images or video frames. By employing image processing, the model becomes more resilient to noise, artifacts, and variations in lighting conditions, resulting in improved detection performance. The proposed deepfakes detection model is trained using a combination of labeled datasets consisting of both authentic and deepfake media samples. The training process involves optimizing the model's

DOI: 10.17148/IJARCCE.2024.131023

parameters using appropriate loss functions, such as binary cross-entropy or adversarial loss, to minimize the discrepancy between predicted and ground truth labels. Training is performed iteratively using backpropagation and gradient descent algorithms to update the model's weights and improve its discriminative capabilities. By combining the deep learning frameworks of ResNeXt and LSTM, along with image processing techniques, the proposed deepfakes detection model offers a powerful solution to combat the dangers posed by deepfake technology. The model's ability to capture both spatial and temporal information enables it to effectively identify manipulated media and mitigate the risks associated with deepfakes, safeguarding individuals and organizations from potential harm.

II. DATASET CREATION

To make the model efficient for real-time prediction, we have gathered data from various available datasets, including FaceForensic++ (FF), Deepfake Detection Challenge (DFDC), and Celeb-DF. The objective is to create a comprehensive dataset that encompasses different types of videos to ensure accurate and real-time detection of deepfake videos. To avoid training bias, the dataset consists of an equal distribution of 50% real videos and 50% fake videos. The Deepfake Detection Challenge (DFDC) dataset includes certain audio-altered videos; however, for the purposes of this paper, audio deepfakes are considered out of scope. Therefore, a python script was used to preprocess the DFDC dataset and remove the audio-altered videos. We selected 1000 real videos and 1000 fake videos from the DFDC dataset. From the FaceForensic++ (FF) dataset, they included 500 real videos and 500 fake videos. Additionally, 500 real videos and 500 fake videos were taken from the Celeb-DF dataset. This combined dataset consists of a total of 2000 real videos, 2000 fake videos, and 4000 videos in total. This dataset is carefully curated to ensure a representative distribution and minimize biases. We preprocess the dataset using image processing techniques, such as resizing, cropping, and normalization, to enhance the quality and consistency of the data.

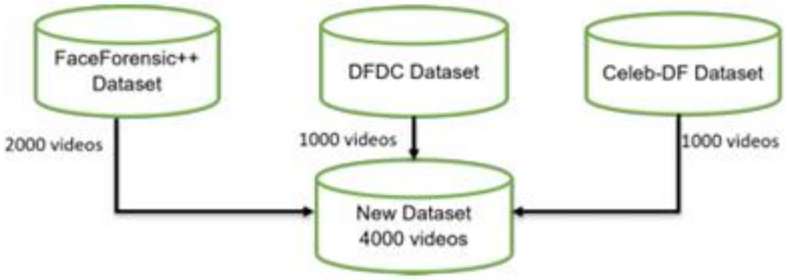


FIG 1 – DEEPFAKE DATASET

III. DATA PREPROCESSING

In the preprocessing stage of the deepfakes detection process, the videos undergo several steps to remove unnecessary noise and focus solely on the face region. Initially, the original video is split into individual frames, with each frame representing a single image from the video. Then, a face detection algorithm or model is applied to detect and locate faces in each frame. Once a face is detected, the frame is cropped to include only the face region, discarding irrelevant information. These cropped frames are then combined to form a new video that exclusively contains the sequence of face-only frames. Frames that do not contain detectable faces are ignored as they do not contribute to the subsequent deepfake detection process. To maintain consistency and handle computational constraints, a threshold value is determined based on the mean total frame count of each video. For example, a threshold of 150 frames may be selected, indicating that only the first 150 frames of each video will be considered for further processing. This threshold is chosen to strike a balance between computational efficiency and accurate analysis. The newly created videos, containing the initial 150 frames, are saved at a standardized frame rate of 30 frames per second (fps) and a resolution of 112 x 112 pixels. This standardization facilitates the proper utilization of the Long Short-Term Memory (LSTM) model, which operates on sequential data. By considering the frames sequentially, the LSTM model can effectively capture temporal dependencies and patterns within the video data, enhancing the deepfake detection process.

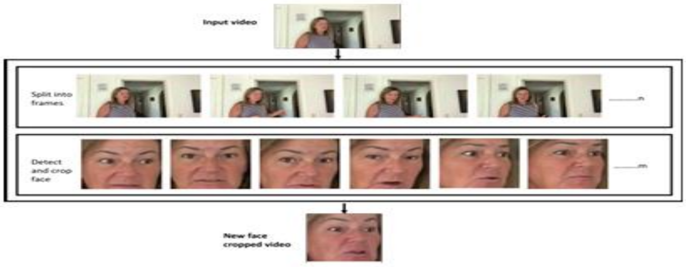


Fig 2 – Preprocessing

By incorporating these additional preprocessing steps, the data is further refined and optimized for deepfake detection. The quality enhancement, face alignment, normalization, data augmentation, and temporal sampling techniques contribute to improved accuracy, robustness, and efficiency in the subsequent stages of deepfake detection.

IV. TRAINING

The deepfake detection system utilizes the ResNext CNN model, specifically the resnext50_32x4d variant, which is a powerful convolutional neural network architecture designed to handle deep networks efficiently. The model consists of 50 layers and has dimensions of 32x4. The Sequential layer is used to store the feature vectors returned by the ResNext model in order, enabling sequential passing of features to the LSTM layer.

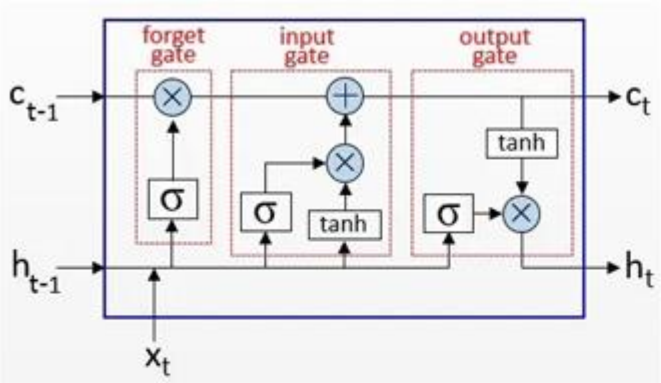


Fig 4 – LSTM architecture

The LSTM layer processes the frames sequentially, capturing temporal changes between video frames by comparing the current frame with earlier frames. It has 2048 latent dimensions and 2048 hidden layers, with a dropout rate of 0.4 to prevent overfitting. The ReLU activation function is employed, providing non-linear behavior and computational efficiency. A dropout layer with a dropout rate of 0.4 is used to enhance the model's robustness and generalization. An Adaptive Average Pooling layer reduces variance, computational complexity, and extracts low-level features from local regions of the feature maps. Together, these components enable the deepfake detection model to effectively process video frames, capture temporal information, and make accurate classifications.

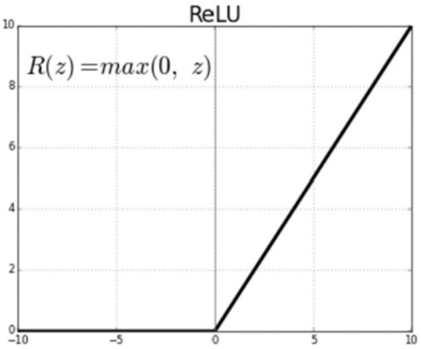


Fig 5 – ReLU activation function

DOI: 10.17148/IJARCCE.2024.131023

In the project, the dataset is divided into a train dataset and a test dataset using a ratio of 70% for the train dataset and 30% for the test dataset. This split ensures that a sufficient amount of data is available for training the deepfake detection model while also reserving a portion for evaluating its performance. Importantly, the split is balanced, meaning that both the train and test datasets contain an equal number of real and fake videos. This balance helps in training a model that can effectively classify both types of videos. To efficiently handle the dataset during training, a data loader is utilized. The data loader is responsible for loading the videos along with their corresponding labels. In this case, a batch size of 4 is chosen, which means that four videos are loaded and processed together in each iteration. This batch processing approach improves training speed and memory utilization, as it allows for parallel computations on multiple videos simultaneously. During the training process, the model undergoes multiple epochs, with each epoch representing a complete iteration through the entire training dataset. In this project, the training process is performed for 20 epochs. The learning rate is set to 1e-5 (0.00001), which determines the step size the optimizer takes in updating the model's parameters. A weight decay of 1e-3 (0.001) is applied to control overfitting by adding a penalty to the loss function based on the magnitude of the model's weights. The Adam optimizer, a popular optimization algorithm in deep learning, is employed to update the model's parameters based on the computed gradients. For the loss function calculation during training, the cross-entropy approach is utilized. Cross entropy is commonly used in classification tasks to measure the dissimilarity between predicted and actual probability distributions. In the context of the deepfake detection model, the cross-entropy loss function quantifies the difference between the predicted probabilities of a video being real or fake and the actual labels.

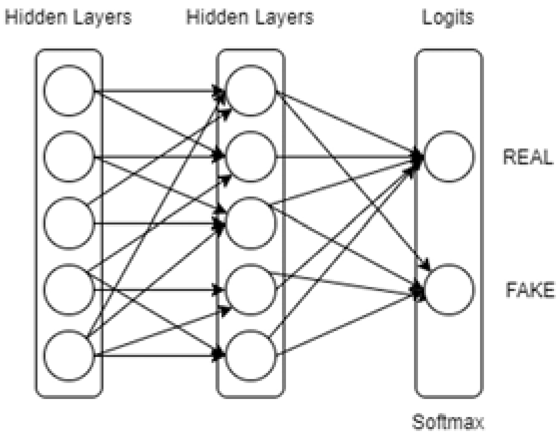


Fig 6 – Softmax layer

As the final layer of the neural network model, a softmax layer is applied. The softmax function is a type of squashing function that normalizes the output of the model to a range between 0 and 1. It is particularly useful in multi-class classification problems, as it provides probabilities for each class prediction. In this project, the softmax layer consists of two output nodes representing the classes "REAL" and "FAKE." The softmax layer's output provides the confidence or probability associated with each prediction, indicating the model's level of certainty in classifying a video as either real or fake.

V. RESULTS

The model performs several operations on the image, including generating feature maps, calculating logits, applying a softmax function, and obtaining a prediction. It also visualizes the prediction by creating a heatmap overlay on the input image. The predict function takes the model and the uploaded image as input. It processes the image through the model and calculates the confidence of the prediction. If the confidence is above a specified threshold, it returns the predicted class and confidence as a list.

DOI: 10.17148/IJARCCE.2024.131023

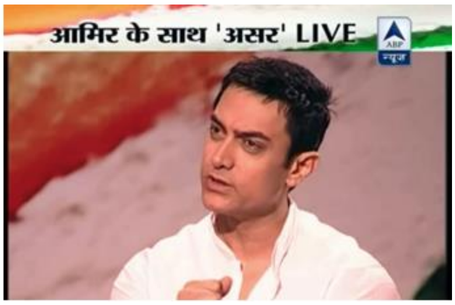
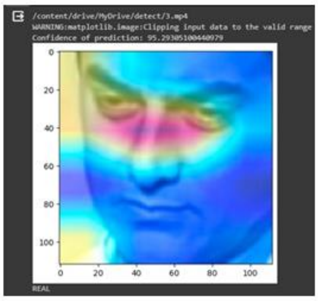


Fig 7 – Input video (Real)



```
/content/drive/MyDrive/detect/3.mp4
WARNING:matplotlib.image:Clipping input data to the valid range
Confidence of prediction: 95.29305100440979
```

```
REAL
```

Fig 8 – Model predection (Real)

Fig 7 and Fig 9 are the inputs to the model and the Fig 8 and Fig 10 are the outputs predicted ny the model. In the above case the first input was real and second was fake input the model predicted correctly. This suggests that the model is capable of distinguishing between real and fake inputs and is making correct predictions in this scenario. The provided code and analysis demonstrate the functionality of the model and its ability to classify images as real or fake based on the provided inputs.

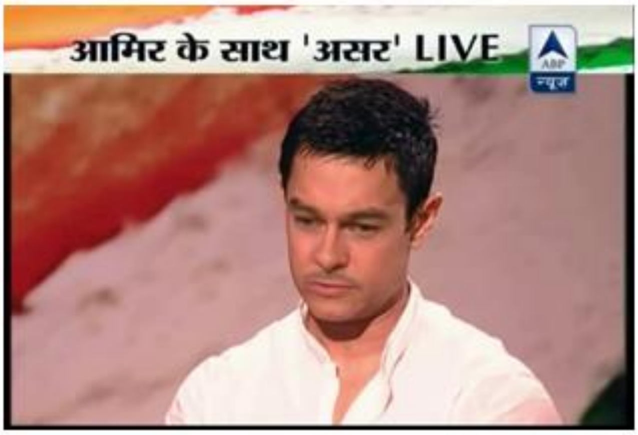
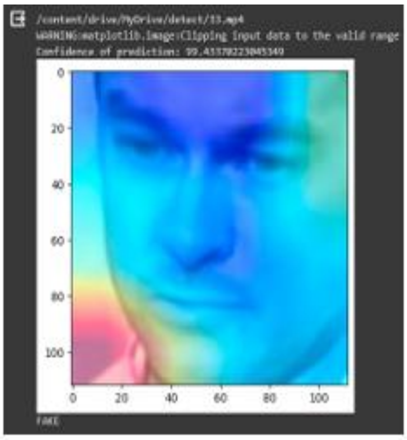


Fig 9 – Input video (Fake)



```
/content/drive/MyDrive/detect/13.mp4
WARNING:matplotlib.image:Clipping input data to the valid range
Confidence of prediction: 99.43378223045349
```

```
FAKE
```

Fig 10 – Model predection (Fake)

In conclusion, the model showcased in the provided code and analysis demonstrates its capability to accurately classify images as real or fake. The correct classification of the real and fake inputs in the given scenario further strengthens the model's reliability and effectiveness. Overall, the demonstrated functionality and accurate predictions highlight the model's potential in image classification tasks, particularly in detecting real and fake inputs.

VI. CONCLUSION

In conclusion, this project aims to develop a deep learning based method for detecting deepfake videos, which pose a significant threat in today's digital landscape. By utilizing a ResNeXt convolutional neural network and an LSTM-based recurrent neural network, the system demonstrates promising results in distinguishing between authentic and manipulated videos. The future work includes the development of a web based platform for users to test their media content for deepfakes, integration into social media platforms for real time fact-checking, and expansion to detect body-based deepfakes. By pursuing these advancements, the project can evolve into a versatile and integrated platform that effectively addresses the challenge of manipulated and synthetic media, promoting trust and authenticity in the digital realm.

REFERENCES

[1]. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). "FaceForensics++: Learning to Detect Manipulated Facial Images." In Proceedings of the IEEE International Conference on Computer Vision (ICCV). C. En Guo, S.-C. Zhu and Y. N. Wu, "Primal Sketch: Integrating Structure and Texture", *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 5-19, 2007.

[2]. Li, Y., Chang, M. C., & Lyu, S. (2018). "In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking." In Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS).

DOI: 10.17148/IJARCCE.2024.131023

[3]. Hsu, C. W., Chang, C. S., Lin, Y. Y., & Wei, S. H. (2018). "A two stream neural network for tampered facial image detection." In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). A. Wierman, Z. Liu, I. Liu and H. Mohsenian-Rad, "Opportunities and challenges for data center demand response", *Proc. Int. Green Comput. Conf.*, vol.7, no. 6, pp.1-10, 2014.

[4]. Bayar, B., & Stamm, M. C. (2016). "A deep learning approach to universal image manipulation detection using a new convolutional layer." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Haoying Dai, Yanne Kouomou Chembo, "RF Fingerprinting Based on Reservoir Computing Using Narrowband Optoelectronic Oscillators", *Journal of Lightwave Technology*, vol.40, no.21, pp.7060-7071, 2022.

[5]. 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). J. Hwang, J. Kim and H. Choi, "A review of magnetic actuation systems and magnetically actuated guidewire-and catheter-based microrobots for vascular interventions", *Intell. Serv. Robot*., vol. 13, no. 1, pp. 1-14, 2020.

[6]. Gazi Hasin Ishrak, Zalish Mahmud, MD. Zami Al Zunaed Farabe, Tahera Khanom Tinni, Tanzim Reza and Mohammad Zavid Parvez “Explainable Deepfake Video Detection using Convolutional Neural Network and CapsuleNet”. B. Accou, J. Vanthornhout, H. V. Hamme and T. Francart, "Decoding of the speech envelope from eeg using the vlaai deep neural network", *Scientific Reports*, vol. 13, no. 1, pp. 812, 2023.

[7]. Aarti Karandikar, Vedita Deshpande, Sanjana Singh, Sayali Nagbhidkar, Saurabh Agrawal.(2020)"Deepfake Video Detection Using Convolutional Neural Network" In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).