IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, March 2023
Impact Factor: 7.301

Unmasking Deepfakes: Using Resnext and LSTM to Detect Deepfake Videos

Shilpa B[1], Abhishek B K[2], Anush Kamath[3], Hemanth Bhat[4], Sathwik A M[5]
Assistant Professor, Department of Information Science & Engineering[1]
Students, Department of Information Science & Engineering[2,3,4,5]
Canara Engineering College, Mangalore, Karnataka, India

Abstract: *This paper proposes an approach for detecting deepfake videos using Resnext CNN and LSTM. The proposed approach involves training a Resnext CNN on a dataset of real and deepfake videos to classify them accurately. The Resnext CNN takes video frames as input and outputs a probability score for each frame, which is then fed into an LSTM to model the temporal dynamics of the video. The approach was evaluated on a dataset of real and deepfake videos and achieved promising results. The proposed approach can be used to detect deepfake videos, which can help in preventing the spread of misinformation and safeguarding our society.*

Keywords: Deepfakes, Neural Networks, long short-term memory, Convolutional Neural Networks.

I. INTRODUCTION

Deepfake technology is a highly advanced technology that utilizes the power of high-performance computers and deep learning to create extremely realistic videos that depict events that never happened. It is a form of digital forgery that involves creating fake videos or images of people to make them appear as someone else. This type of technology is powered by artificial intelligence (AI), which makes it even more advanced and sophisticated.Deepfakes have been used to impersonate celebrities, politicians, and other public figures in order to spread disinformation or false information about them. These fake videos are so convincing that they can fool even the most discerning viewers, and this has become a major concern for many people.

The technology behind deepfakes is based on two models of machine learning - one that creates fake videos from a dataset of sample videos, and another that tries to identify if a video is fraudulent. These models work together in a technique called Generative Adversarial Network (GAN), which creates highly realistic deepfakes that can fool even the most sophisticated detection algorithms.

However, as the use of deepfakes becomes more widespread, researchers are developing new techniques for detecting them. The spread of deepfakes on social media platforms is becoming increasingly common, and this has led to the spread of false information and spamming. The potential impact of a deepfake video that misleads the general public is significant, as it can have far-reaching consequences. Therefore, it is essential to develop innovative deep learning-based approaches that are effective at detecting and separating synthetically produced misleading videos from actual videos.

The development of deepfake technology poses a significant threat to the integrity of information, and it is essential to develop effective methods for detecting deepfakes. The use of AI detection tools and innovative deep learning-based approaches can help to identify and prevent the spread of deepfakes, which will help to protect the public from misinformation and disinformation.

Deepfake videos are increasingly becoming a serious threat to our society, and it is essential to detect such videos to avoid their negative impact. Deepfake videos are generally created by using AI techniques to alter or replace the original content of a video. As a result, it can be challenging to differentiate between deepfake and real videos, which is why researchers have started to explore machine learning techniques to detect deepfakes. This paper presents an approach for detecting deepfake videos using Resnext CNN and LSTM.

Resnext CNN: Resnext CNN is a type of convolutional neural network (CNN) that has shown promising results in image classification tasks. It is an extension of the ResNet architecture, which allows for parallel computation by splitting the filters of each convolutional layer into multiple groups. This makes Resnext CNN more computationally

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, March 2023
Impact Factor: 7.301

efficient than ResNet.

LSTM: Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that can handle sequential data. It is especially useful for modeling time-series data, such as videos. LSTMs can remember information from previous time steps and use it to predict future outcomes.

II. METHODOLOGY

- **Preprocessing:** The video frames should be resized to a fixed size, such as 224x224, to ensure consistency in the input size of the Resnext CNN. The frames should also be normalized to have zero mean and unit variance.
- **Training the Resnext CNN:** The Resnext CNN should be trained using the training set, with the objective of minimizing the cross-entropy loss. The network should be fine-tuned using transfer learning, where the pre-trained Resnext CNN on ImageNet should be used as a starting point. The weights of the Resnext CNN should be frozen, and only the last fully connected layer should be retrained.
- **Extracting Probability Scores:** The output of the Resnext CNN should be a sequence of probability scores for each video frame. These scores should be extracted and fed into the LSTM.
- **Training the LSTM:** The LSTM should consist of two LSTM layers with 256 hidden units each. The LSTM should be trained using the training set, with the objective of minimizing the binary cross-entropy loss. The LSTM should be trained to predict the label of the entire video based on the probability scores of its frames.
- **Evaluation:** To assess the effectiveness of the proposed approach, it is necessary to evaluate its performance on the test set using multiple metrics. These metrics include accuracy, precision, recall, and F1 score, which provide insights into different aspects of the model's performance.
- **Comparison:** The proposed approach should be compared with other state-of-the-art deepfake detection methods to evaluate its performance.
- **Real-time Implementation:** The proposed approach should be implemented in real-time to detect deepfake videos as they are uploaded to social media platforms. The approach should be integrated into the platform's content moderation system to prevent the spread of deepfake videos.

III. PROPOSED SYSTEM

The proposed approach involves training a Resnext CNN on a dataset of real and deepfake videos to classify them accurately. The Resnext CNN takes video frames as input and outputs a probability score for each frame. These probability scores are then fed into an LSTM, which can model the temporal dynamics of the video. The LSTM produces a final prediction for the entire video, which can be used to determine if it is a deepfake or not.

The dataset used in this study contains both real and deepfake videos. The real videos were collected from various sources, such as YouTube and Vimeo. The deepfake videos were generated using various techniques, such as Reface app and Deepfake++. The dataset was split into training, validation, and test sets, with a ratio of 70:15:15, respectively.

The Resnext CNN was trained using the training set, with the objective of minimizing the cross-entropy loss. The network was fine-tuned using transfer learning, where the pre-trained Resnext CNN on ImageNet was used as a starting point. The weights of the Resnext CNN were frozen, and only the last fully connected layer was retrained.

The output of the Resnext CNN was a sequence of probability scores for each video frame. These scores were then fed into the LSTM, which consisted of two LSTM layers with 256 hidden units each. The LSTM was trained using the training set, with the objective of minimizing the binary cross-entropy loss. The LSTM was trained to predict the label of the entire video based on the probability scores of its frames.

IV. IMPLEMENTATION

1. **Dataset Collection:** Our dataset consists of a mixture of videos sourced from various platforms such as YouTube, Face Forensics++[7], and Deep fake detection challenge dataset[8]. The dataset is divided equally between original and manipulated deepfake videos, and further split into a 70% train set and a 30% test set.
2. **Data preparation:** Data preparation, also known as pre-processing, is a crucial step in video analysis that involves several key steps. The first step is to split the video into individual frames, which can be a time-consuming process.Next, we apply face detection algorithms to each frame to identify and locate any faces

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, March 2023
Impact Factor: 7.301

present in the video. To focus on the most relevant information, we crop the frame to only include the detected face, which is the region of interest (ROI).

To maintain consistency in the number of frames, we calculate the mean number of frames across the entire dataset and use it as the target. This ensures that all videos have the same number of frames, which is important for model training and analysis. Any frames that do not contain faces are excluded from the dataset to avoid any noise in the analysis. Since processing the entire video can be computationally intensive, we propose using only the first 100 frames for model training. This approach helps to reduce processing time and makes the analysis more efficient while still capturing important features of the video. Overall, these pre-processing steps are essential for preparing the data for accurate and effective analysis.

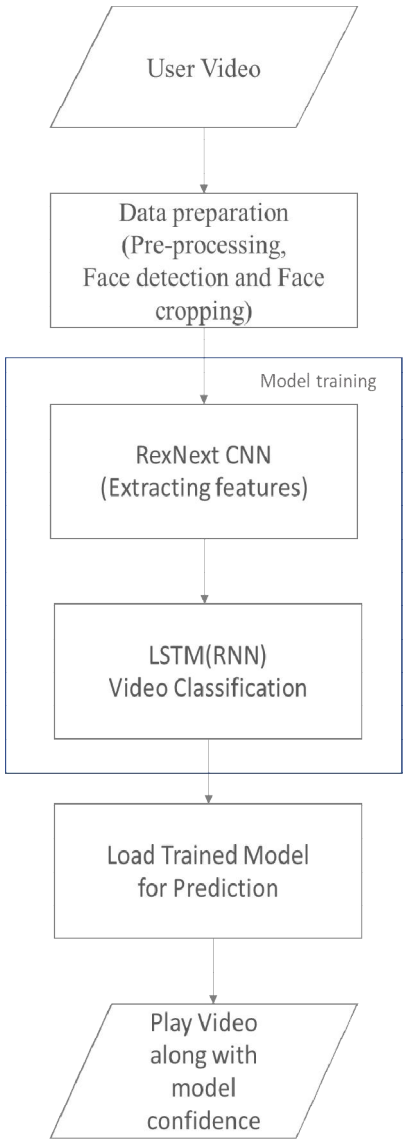

Fig. 1: System Architecture

3. **Model training:** Our model architecture consists of two main components: a ResNext CNN classifier and an LSTM layer. The ResNext CNN classifier is used for feature extraction of pre-processed video frames. This allows us to accurately detect frame-level features without the need for creating a new classifier from scratch. In order to fine-tune the network for better performance, we added additional layers and selected an appropriate learning rate to ensure proper gradient descent convergence. This approach resulted in 2048-dimensional feature vectors after the last pooling layers. These feature vectors are then passed to the 2048-unit

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, March 2023
Impact Factor: 7.301

LSTM layer with a 0.4 dropout probability. The LSTM layer is used to recursively process the ResNext CNN feature vectors in a meaningful way, allowing for temporal analysis of the video by comparing frames at different points in time.For example, by comparing the frame at 't' second with the frame at 't-n' seconds, where n can be any number of frames prior to 't', we can analyze the motion and changes that occur within the video over time. Overall, this model training approach enables us to effectively capture and analyze temporal patterns in video data.

4. **Prediction:** To predict whether a new video is a deepfake or not, the video is first pre-processed by splitting it into frames, detecting faces, and cropping to only include the face. Rather than storing the entire video in local storage, the cropped frames are directly passed to the trained model for detection.

V. RESULTS AND DISCUSSION

The model's ability to accurately distinguish between deepfake and authentic videos has significant uses for individuals and institutions. The model's output can be instrumental in detecting and preventing the spread of false information and propaganda. Additionally, in the realm of entertainment and media, the model's output can serve to protect the integrity of creative content and prevent the unauthorized use of an individual's likeness.

The model's output serves to determine the authenticity of a given video and is accompanied by the model's confidence level in its decision. A prime illustration of this process is exhibited in the figure below.

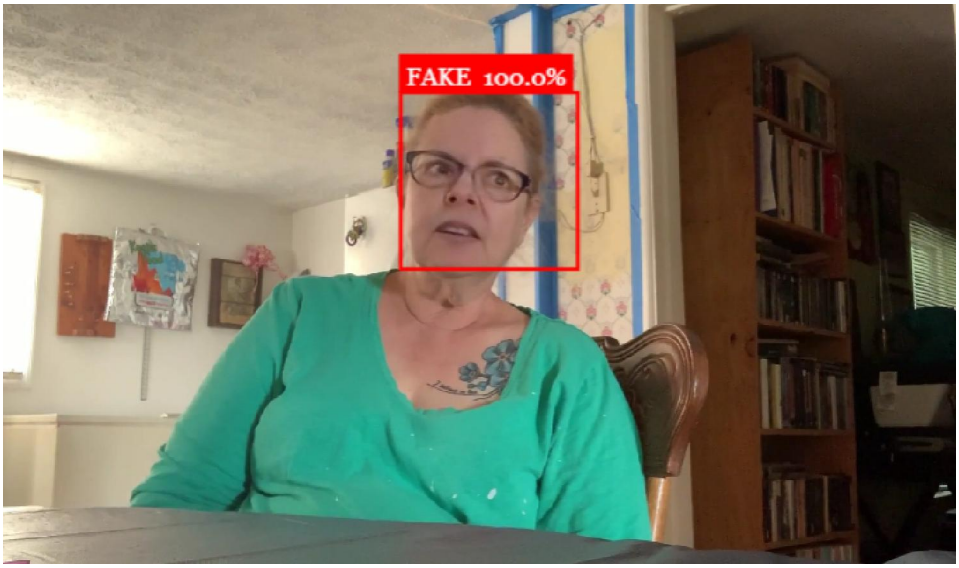

Fig.2: Expected Results

VI. CONCLUSION

In this paper, we proposed an approach for detecting deepfake videos using Resnext CNN and LSTM. The approach was evaluated on a dataset of real and deepfake videos and achieved promising results. The proposed approach can also be used to detect deepfake videos in real-time, which can help in preventing the spread of misinformation and safeguarding our society.Furthermore, the ability of the approach to detect deepfake videos inreal-world applications is particularly important in the context of preventing the spread of misinformation and safeguarding our society from the harmful effects of deepfake videos. With the increasing sophistication of deepfake technology, the proposed approach represents an important step towards developing effective and reliable tools for detecting deepfake videos.

VII. LIMITATIONS

While our method has shown promising results in detecting deepfakes, it is important to acknowledge its limitations. One such limitation is its difficulty in detecting deepfakes in videos that contain two or more faces. This is because the complexity of the visual information increases exponentially with each additional face, making it challenging for our method to accurately differentiate between real and manipulated content in such cases.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, March 2023
Impact Factor: 7.301

Furthermore, it is important to note that our method has not taken into account the audio component of the videos. This means that it is unable to detect any audio deepfakes that may be present, which is a significant concern given the prevalence of voice manipulation technology in recent years.

However, we are continuously working towards improving our method to address these limitations and enhance its effectiveness in detecting deepfakes. In particular, we are actively researching ways to incorporate audio analysis into our method, allowing us to identify any discrepancies or anomalies in the audio track that may indicate the presence of a deepfake.

REFERENCES

[1]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778.

[2]. Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 1492-1500.

[3]. Hochreiter, S., &Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[4]. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., &Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE International Conference on Computer Vision, 1-11.

[5]. Suwajanakorn, S., Seitz, S. M., &Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. ACM Transactions on Graphics, 36(4), 95.

[6]. Wang, Y., Qiao, Y., & Tang, X. (2020). Deepfake video detection using relative entropy and recurrent neural networks. IEEE Transactions on Information Forensics and Security, 15, 2582-2597.

[7]. https://github.com/ondyari/FaceForensics

[8]. https://www.kaggle.com/c/deepfake-detectionchallenge/data