,,

A Mini Project Report on

Vigilens: Unmasking DeepFake Videos Using Deep Learning

Submitted in partial fulfilment for the degree of Bachelor of Technology in Computer Science and Technology

Submitted by

Trithi Amin (211002)
Sanskruti Bagul (211005)
Anusha Goyal (211017)

Under the guidance of

Prof. Monica Charate

USHA MITTAL INSTITUTE OF TECHNOLOGY
S.N.D.T WOMEN'S UNIVERSITY
MUMBAI – 400049

2023 – 2024

CERTIFICATE

This is to certify that **Ms. Trithi Amin, Ms. Sanskruti Bagul, Ms. Anusha Goyal** has successfully completed minor project work on **Vigilens: Unmasking DeepFake Videos Using DeepLearning** in the partial fulfillment for the bachelor's degree in **Computer Science and Technology** during the year 2023-24 as prescribed by SNDT Women's University.

GUIDE
PROF. MONICA CHARATE

H.O.D
PROF. KUMUD WASNIK

PRINCIPAL
DR. YOGESH NERKAR

EXAMINAR 1

EXAMINAR 2

ACKNOWLEDGEMENT

We would like to express our sincere appreciation to everyone who gave their valuable assistance and support in this project. We are highly indebted to Prof. Monica Charate for her guidance and constant supervision as well as for providing necessary information and support in this project. Without her help, this project would not have been possible.

We would also like to Thank our Head of Department Prof. Kumud Wasnik for her insights in shaping the direction and content of this report. We would also like to extend our gratitude to our classmates who provided valuable feedback and encouragement throughout the process.

Finally, we would like to express our appreciation to our families for their unwavering support and understanding during the long hours and late nights spent working on this project.

Thank you all for your contributions, support, and encouragement.

ABSTRACT

The project titled "Vigilens: Unmasking DeepFake Videos Using Deep Learning" is a comprehensive study on detecting deepfake videos using deep learning techniques. The project aims to address the growing concern of deepfake technology, which has become increasingly sophisticated due to the advancement of deep learning algorithms. The project's core lies in the implementation of a deep learning model that can recognize specific patterns and features associated with unmasked deep fakes.

The project uses Convolutional Neural Networks (CNN) to extract frame-level features and a Long Short Term Memory (LSTM) based Recurrent Neural Network (RNN) to classify whether the video is a deep fake or real video. The model is trained and evaluated on a large amount of balanced and mixed data-set prepared by mixing various available data-sets. The methodology involves constant refinement to adapt to emerging deepfake techniques and variations.

The project's primary objective is to develop a vigilant system that can distinguish between authentic and manipulated media. The system not only aims to detect manipulated content but also to shed light on the potential ethical impacts and consequences of unbridled deepfake proliferation. The project seeks to provide insights into countering the negative implications while fostering responsible development and usage of synthetic media technology.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Background

With the rapid advancement of deep learning techniques, particularly in the realm of generative adversarial networks (GANs), the creation of convincing deepfake videos has become increasingly prevalent. Deepfake technology allows for the manipulation of audio and video content, often leading to the creation of misleading or harmful media content. As such, there is a pressing need for robust deepfake detection systems to combat the spread of misinformation.

1.2 Purpose of the object

The purpose of the proposed deepfake detection system is to leverage aforementioned techniques in Recurrent neural networks (RNN), long short-term memory (LSTM) networks to accurately identify and classify deepfake media. By integrating RestnextCNN, a variant of CNN known for its efficiency and effectiveness in image recognition tasks, with LSTM, a type of recurrent neural network capable of capturing temporal dependencies in sequential data, the system aims to analyse both spatial and temporal features inherent in deepfake videos. Through this integration, the system seeks to enhance its ability to discern subtle discrepancies between authentic and manipulated media content.

1.3 Objective of the project

We aim to develop models tailored for bigtech companies to seamlessly integrate our backend model into comprehensive applications. Our focus lies in crafting a deepfake detection system that stands as a dependable and scalable solution across varied platforms and applications. Our approach involves training the model on diverse datasets encompassing both authentic and deepfake media samples, facilitating the acquisition of discerning patterns to differentiate between genuine and manipulated content. Moreover, we emphasize the paramount importance of efficiency and performance, ensuring rapid and precise identification of deepfakes. Ultimately, our overarching goal is to counteract the adverse effects of deepfake technology.

1.4 Problem Statement

Convincing manipulations of digital images and videos have been demonstrated for several decades through the use of visual effects, recent advances in deep learning have led to a dramatic increase in the realism of fake content and the accessibility in which it can be created. These so-called AI-synthesised media (popularly referred to as deep fakes). Creating the Deep Fakes using the Artificially intelligent tools is a simple task. But, when it comes to detection of these Deep Fakes, it is a major challenge. Already in the history there are many examples where the deepfakes are used as powerful way to create political tension, fake terrorism events, revenge porn, blackmail peoples etc. So it becomes very important to detect these deepfake and avoid the percolation of deepfake through social media platforms.

This project aims to distinguish and classify the video as deepfake or pristine using LSTM based artificial Neural networks.

Chapter 2

Literature Survey

2.1 Technical Papers

A literature survey, also known as a literature review, is an essential part of research that involves a comprehensive analysis and synthesis of previously published research on a specific topic. The purpose of a literature survey is to identify gaps in knowledge, assess the state of research, and provide a foundation for further research.

Throughout the research phase, we looked on to various research papers and articles citing about deepfake detection system. Amongst all we selected 5 papers for the successfull completion of this project. The Research Paper No. 1 is the base of this project, as it tells the challenges and next steps to be taken to overcome those challenges. All the research papers were very insightful and helped us in the proper understanding of the problem and creating an effective solution for it. With the help of research papers we were able to design the architecture of our model.

The following table provides an overview of this Literature survey:

Sr. No.	Paper Name, Author(s), Year of Publishment	Methodology and Technologies	Observations/ Findings and Remarks
1.	Adversarially Robust DeepFake Media Detection Using Fused Convolutional Neural Network Predictions. Authors : Sohail Ahmed Khan , Dr. Alessandro Artusi , Dr. Hang Dai Yr: 2021	The paper addresses the challenge of deepfake detection systems struggling against unseen data by employing three different deep CNN models, to classify fake and real images extracted from videos. The document emphasizes the importance of image pre-processing and augmentation techniques in enhancing deepfake detection, along with the significance of robustness against adversarial attacks.	The proposed technique outperforms state-of-the-art models, achieving 96.5 percent accuracy on the publicly available DeepFake Detection Challenge (DFDC) test data. Additionally, the fusion model achieves 99 percent accuracy on lower quality DeepFake-TIMIT dataset videos and 91.88 percent on higher quality DeepFakeTIMIT videos.
2.	Towards Solving the DeepFake Problem: An Analysis on Improving DeepFake Detection Using Dynamic Face Augmentation. Authors: Sowmen Das , Selim Seferbekov , Arup Datta , Md. Saiful Islam , Md. Ruhul Amin Yr: 2018	The research paper discusses various methodologies aimed at enhancing Deepfake detection accuracy. Here are the key methodologies outlined in the paper: Dynamic Face Augmentation, Identifying Dataset Issues, Face Clustering, Pre-processing Guidelines, Experimental Setup and Results	The analysis conducted in the study offers substantial insights into the importance of dataset quality and augmentation methods in enhancing Deepfake detection accuracy and model generalisation. Creating an information fusion model and gathering different types of information is challenging.
3.	Swapped face detection using deep learning and subjective assessment. Authors:Xinyi Ding, Zohreh Raziei, Eric C. Larson, Eli V. Olinick, Paul Krueger and Michael Hahsler Yr: 2020	Development of a deep learning model using transfer learning for detecting swapped faces, a technique often used for deceptive purposes providing high accuracy predictions coupled with an analysis of uncertainties.” They created a dataset comprising over 420,000 images derived from pictures of 86 celebrities.	The study concludes by emphasising the effectiveness of their deep learning model for detecting swapped faces. Model seemed to struggle against higher resolution deepfake videos .

4.	DeepFake Detection: Current Challenges and Next Steps. Author : Siwei Lyu Yr: 2020	The paper discusses the emergence of deepfake videos, including head puppetry, face swapping, and lip syncing. It discusses the significant progress in effective detection method including large scale deepfake video datasets and public challenges dedicated to deepfake detection.	The classification success rate was increased with the training of the respective networks. It provides an overview of future technological developments in terms of running efficiency , detection efficiency,accuracy and robustness.
5.	Robust Face-Swap Detection Based on 3D Facial Shape Information. Authors: Weinan Guan, Wei Wang, Jing Dong, Bo Peng , Tieniu Tan Yr: 2021	To capture the inconsistency of 3D facial shape in face-swap images and videos,they utilised 3DMM (3D morphable model) to extract 3D facial shape features of face-swap images and videos. They calculate Mahalanobis Distance between the shape features of suspected images and corresponding templates, distance is further utilised to authenticate the suspected images by comparing with the fixed threshold.	Approach is less vulnerable to laundering counter- measures and has good robustness against unseen face-swap methods. 3D facial shape information plays a crucial role to detect face-swap images

Table 2.1: All the Referred Research Papers

2.2 EXISTING SYSTEM

The deepfake detection system is to provide a reliable and scalable solution for identifying deepfake videos across various platforms and applications. There are existing systems of deepfake detection that use various methods to detect the signs of faking. Some of the existing methods and technologies used in deepfake detection systems include:

1. Data-Augmentation System: Data augmentation technique also called Face-Cutout, which dynamically generates training images with occlusions,referring to the act of covering or blocking part of an image. By dynamically adding occlusions to training images, the neural networks are encouraged to focus on learning to detect manipulation features in deepfake content rather than getting fixated on recognizing individual faces.

2. Fused CNN Prediction System: Employing a fusion of deep CNN models in deepfake detection significantly improves the robustness and generalization capability of the detection system. The fusion model, which combines predictions from multiple deep CNN models outperforms individual models. Additionally, the fusion model demonstrates resilience against adversarial attacks, ensuring that if one model is compromised, it does not affect the overall classification.

3. Transfer Learning System: In transfer learning, a machine exploits the knowledge gained from a previous task to improve generalization about another.Learning process can be more accurate and faster or need less training data. Using transfer learning for detecting swapped faces, achieved high accuracy predictions coupled with an analysis of uncertainties.

Chapter 3

PROPOSED SYSTEM

The deepfake video detection model project is structured into several major steps, each crucial for the successful development and deployment of the system. The details of each step are outlined below:

Step 1: Data-set Gathering and Analysis This step involves downloading the dataset relevant to deepfake videos, performing an in-depth analysis of the dataset, and preparing it for preprocessing.

Step 2: Module 1 Implementation Module 1 implementation focuses on splitting the video into frames and cropping each frame to extract the face, a critical step in identifying potential deepfake manipulations.

Step 3: Pre-processing Pre-processing includes creating a new dataset that exclusively contains face-cropped videos, which are essential for further analysis and model training.

Step 4: Module 2 Implementation Module 2 implementation comprises developing a DataLoader for efficiently loading videos and labels, as well as training a baseline model on a small dataset to establish initial performance benchmarks.

Step 5: Hyperparameter Tuning This step involves iteratively adjusting hyperparameters such as learning rate, batch size, weight decay, and model architecture to optimize the model's accuracy until reaching the maximum achievable accuracy.

Step 6: Training the Final Model The final model is trained on a large dataset using the best hyperparameters identified in Step 5, ensuring optimal performance and robustness against deepfake manipulations.

3.1 Workflow of Project

A test video is fed into the pre-processing algorithm for detection of frames, extraction and cropping of faces. Then the processed video is fed into the trained model which will predict whether the video is deepfake or real.

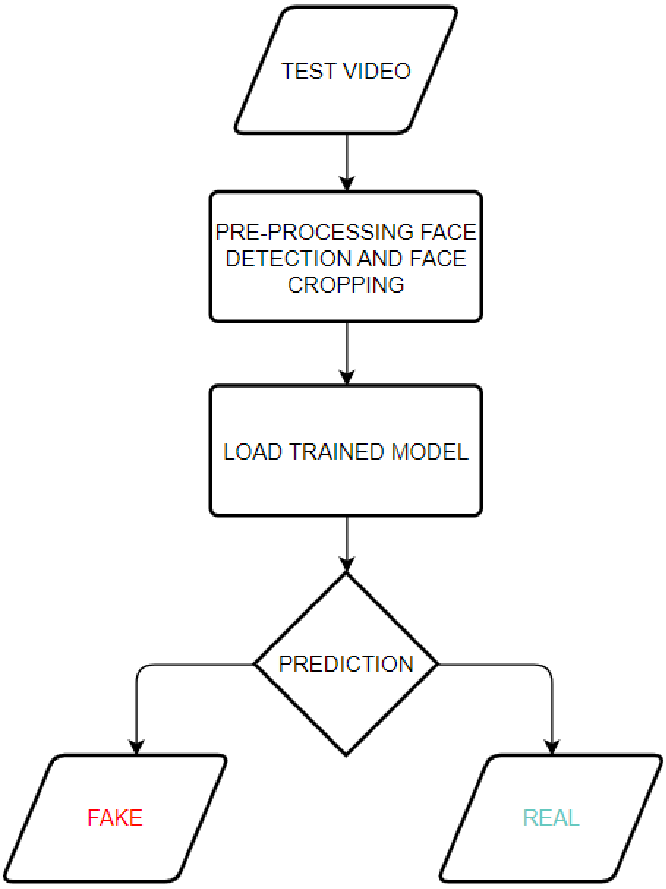

Figure 3.1: Prediction Workflow

Chapter 4

ARCHITECTURE OVERVIEW

The deepfake detection model architecture combines ResNextCNN for spatial feature extraction and LSTM networks for temporal aggregation. ResNextCNN captures high-level spatial features from video frames, while LSTM units analyze sequential data to detect temporal patterns across frames. By integrating these components, the model effectively identifies subtle inconsistencies introduced by deepfake manipulation, enhancing the accuracy and robustness of deepfake detection systems.

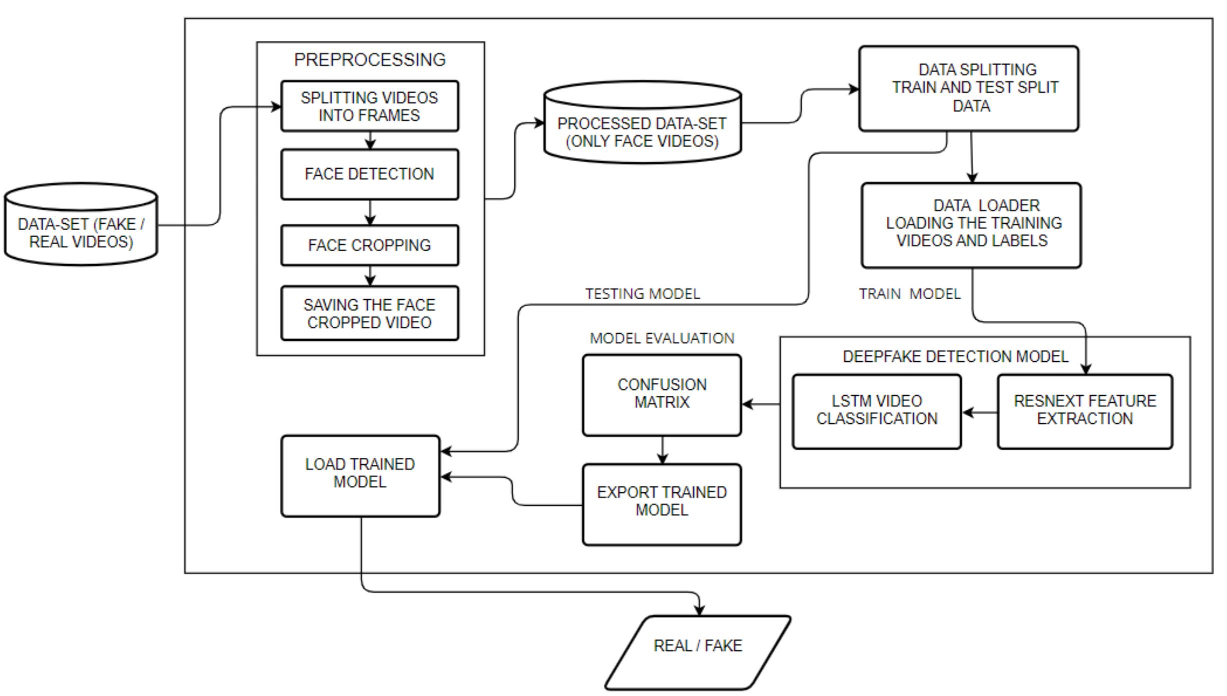

Figure 4.1: Architecture

Chapter 5

HARDWARE AND SOFTWARE REQUIREMENT

5.1 Hardware Specification

- System Processor: Intel Xeon E5 2637: 3.5GHz.
- RAM: Minimum 16 GB
- Hard disk: Minimum 100 GB hard disk space.
- Graphics Card: NVIDIA GeForce GTX Titan

5.2 Software Requirements

- Operating System : Windows 7 or above
- Programming language: Python 3.0
- Framework: PyTorch 1.4
- Libraries: OpenCV, torch, torchvision, os, numpy, cv2, matplotlib, face recognition, json, pandas, copy, glob, random, sklearn
- Tools: Google Colab

Chapter 6

IMPLEMENTATION OF PROJECT

6.1 Dataset Procurement

To gather a comprehensive dataset for training and testing our deepfake detection system, we conducted extensive research across various resources specializing in deepfake datasets. The primary objective was to acquire a dataset that encompasses diverse facial expressions, lighting conditions, and scenarios commonly encountered in real-world applications.

6.2 Sampling Strategy

Define Sampling: The sampling technique employed is purposive sampling. This method was chosen to ensure the selection of datasets that encompass diverse facial expressions, lighting conditions, and scenarios commonly encountered in real-world applications. Additionally, the Celeb Deepfake (Celeb DF) datasets version 2 were augmented with the inclusion of a face forensic dataset to enhance the diversity and representativeness of the sample.

Sample Size: The sample size of Celeb DF version 2, supplemented by the face forensic dataset, was chosen to provide a substantial and varied representation of deepfake instances. This sample size was deemed sufficient to support robust model training and evaluation.

6.3 Data Collection Methods

Detail Data Collection: Data was gathered through extensive research across various resources specializing in deepfake datasets. The primary objective was to identify datasets that meet the criteria of diversity in facial expressions, lighting conditions, and scenarios. Specifically, the Celeb Deepfake dataset v2 was augmented with a face forensic dataset to enrich the sample.

Instruments Used: The tools and instruments utilized for data collection included online repositories, research publications, and specialized platforms hosting deepfake datasets.

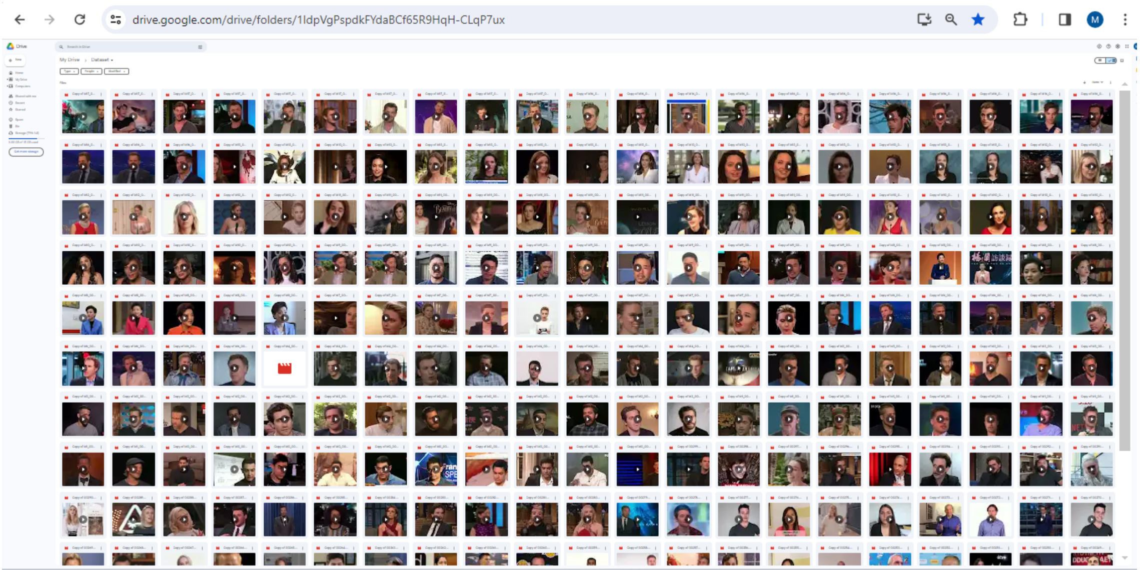

Figure 6.1: Snippet of Celeb VF2 Dataset

6.4 Pre-Processing Details

- Using `glob`, we imported all the videos in the directory into a Python list.
- `cv2.VideoCapture` is used to read the videos and get the mean number of frames in each video. To maintain uniformity, a value of 150 frames is selected as the ideal value for creating the new dataset.
- The video is split into frames, and the frames are cropped to the face location.
- The cropped face frames are then written to a new video using `VideoWriter`.
- The new video is written at 30 frames per second and with a resolution of 112 x 112 pixels in the mp4 format.
- Instead of selecting random videos, the first 150 frames are used to create the new video, ensuring proper use of LSTM for temporal sequence analysis.

6.5 Model Details

Our model is a combination of CNN and RNN. We have used the Pre-trained ResNext CNN model to extract the features at the frame level, and based on the extracted features, an LSTM network is trained to classify the video as deepfake or pristine. Using the Data Loader on the 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 a Residual CNN network optimized for high performance on deeper neural networks. For experimental purposes, we used the `resnext50_32x4d` model, which has 50 layers and dimensions of 32 x 4.

Following this, we will fine-tune 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 layer of ResNext are used as the sequential LSTM input.

LSTM for Sequence Processing: The 2048-dimensional feature vectors are fitted as the input to the LSTM. We are using 1 LSTM layer with 2048 latent dimensions and 2048 hidden layers along with a 0.4 chance of dropout, which is capable of achieving our objective. LSTM is used to process the frames sequentially 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. The model also consists of a Leaky ReLU activation function. A linear layer of 2048 input features and 2 output features is used to make the model capable of learning the average rate of correlation between the input and output. An adaptive average pooling layer with the output parameter 1 is used in the model, which gives the target output size of the image in the form H x W. For sequential processing of the frames, a Sequential Layer is used. The batch size of 4 is used to perform the batch training. A SoftMax layer is used to get the confidence of the model during prediction.

The model consists of the following layers:

- **ResNext CNN:** The pre-trained model of Residual Convolution Neural Network is used. The model name is `resnext50_32x4d`. This model consists of 50 layers and dimensions of 32 x 4.

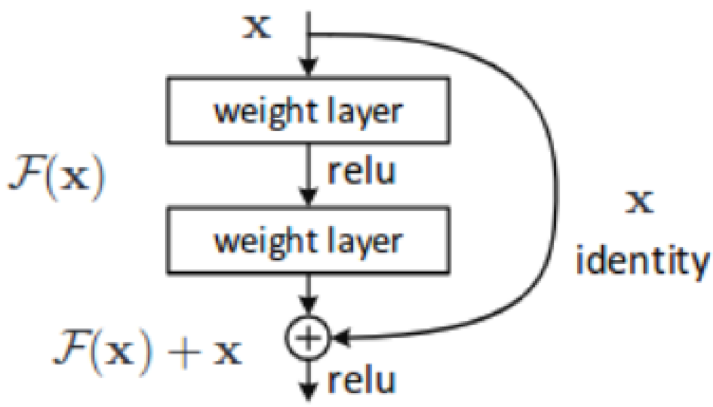

Figure 6.2: ResNext Working

stage	output	**ResNeXt-50 (32×4d)**
conv1	112×112	7×7, 64, stride 2
conv2	56×56	3×3 max pool, stride 2 $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128, C{=}32 \\ 1\times1, 256 \end{bmatrix} \times 3$
conv3	28×28	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256, C{=}32 \\ 1\times1, 512 \end{bmatrix} \times 4$
conv4	14×14	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512, C{=}32 \\ 1\times1, 1024 \end{bmatrix} \times 6$
conv5	7×7	$\begin{bmatrix} 1\times1, 1024 \\ 3\times3, 1024, C{=}32 \\ 1\times1, 2048 \end{bmatrix} \times 3$
	1×1	global average pool 1000-d fc, softmax
# params.		$\mathbf{25.0}\times10^6$

Figure 6.3: ResNext Architecture

- **Sequential Layer:** Sequential is a container of Modules that can be stacked together and run at the same time. The Sequential layer is used to store the feature vector returned by the ResNext model in an ordered way so that it can be passed to the LSTM sequentially.

- **LSTM Layer:** LSTM is used for sequence processing to spot the temporal change between frames. The 2048-dimensional feature vectors are fitted as the input to the LSTM. We are using 1 LSTM layer with 2048 latent dimensions and 2048 hidden layers along with a 0.4 chance of dropout, which is capable of achieving 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.

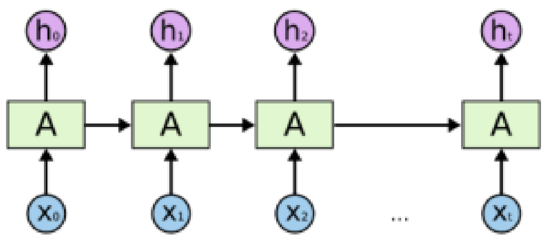

Figure 6.4: Overview of LSTM Architecture

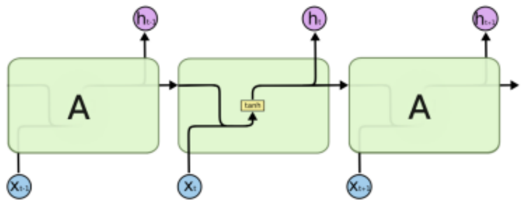

Figure 6.5: Internal LSTM Architecture

- **ReLU:** A Rectified Linear Unit (ReLU) is an activation function that outputs 0 if the input is less than 0, and the raw input otherwise. ReLU is non-linear and has the advantage of not introducing any backpropagation errors, unlike the sigmoid function. It also allows for faster model building, especially for larger Neural Networks.

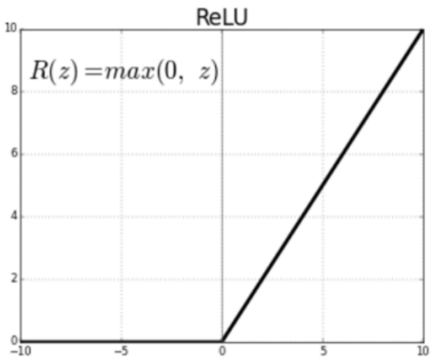

Figure 6.6: ReLU

- **Adaptive Average Pooling Layer:** The Adaptive Average Pooling Layer is used to reduce variance, computation complexity, and extract low-level features from the neighborhood. A 2-dimensional Adaptive Average Pooling Layer is used in the model.

6.6 Model Training Details

- **Train Test Split:** The dataset is split into train and test datasets with a ratio of 70
- **Data Loader:** It is used to load the videos and their labels with a batch size of 4.
- **Training:** The training is done for 20 epochs with a learning rate of 1e-5 (0.00001) and weight decay of 1e-3 (0.001) using the Adam optimizer.
- **Adam optimizer:** To enable the adaptive learning rate, the Adam optimizer with the model parameters is used.
- **Cross Entropy:** To calculate the loss function, the Cross Entropy approach is used because we are training a classification problem.

- **Softmax Layer:** A Softmax function is a type of squashing function that limits the output to the range 0 to 1, allowing it to be interpreted as a probability. It is typically the final layer used in neural network functions for multi-class classification, providing the confidence (probability) of prediction. In our case, the softmax layer has two output nodes: REAL or FAKE.

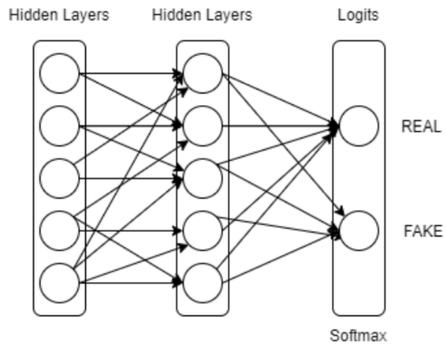

Figure 6.7: Softmax Layer

- **Confusion Matrix:** A confusion matrix is a summary of prediction results on a classification problem. It shows the ways in which your classification model is confused when making predictions, providing insights into the types of errors made. It is used to evaluate our model and calculate accuracy.
- **Export Model:** After training, the model is exported for use in predicting real-time data.

6.7 Model Prediction Details

- The new video for prediction is preprocessed and passed to the loaded model for prediction.
- The trained model performs the prediction and returns whether the video is real or fake, along with the confidence of the prediction.

6.8 Results and Output

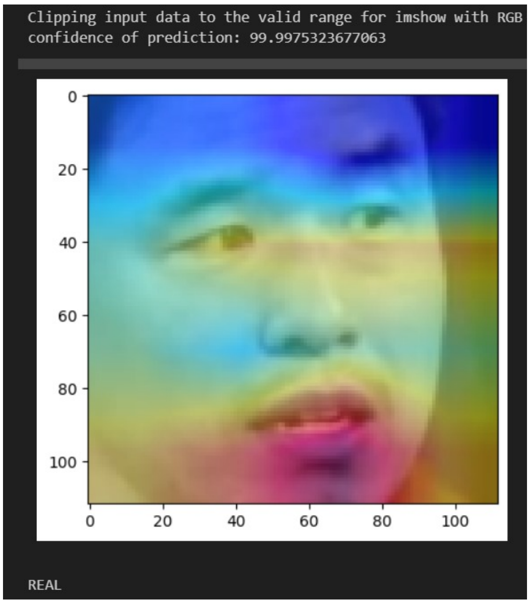

Figure 6.8: Real Output

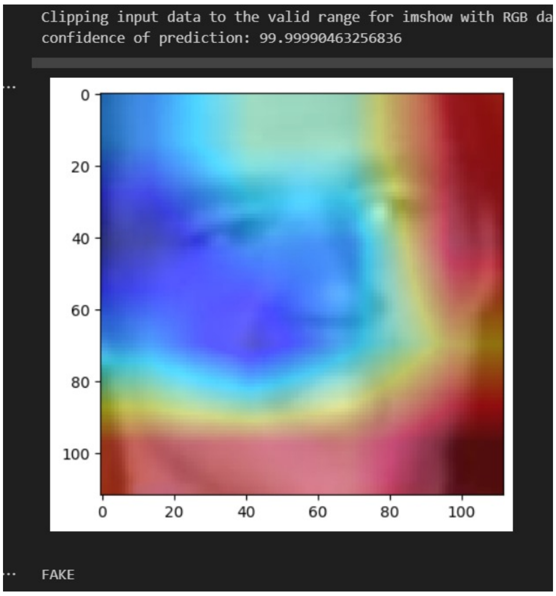

Figure 6.9: Fake Output

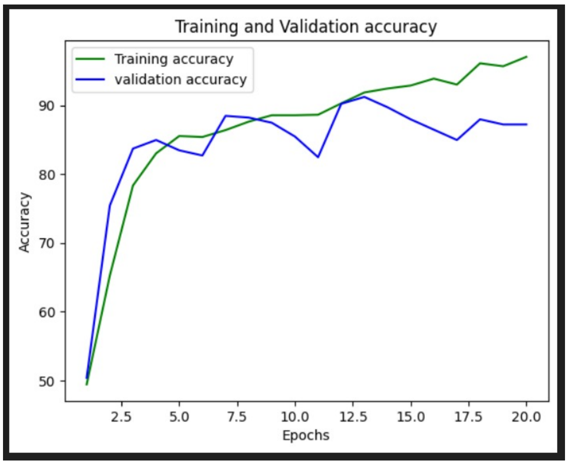

Figure 6.10: Training Accuracy

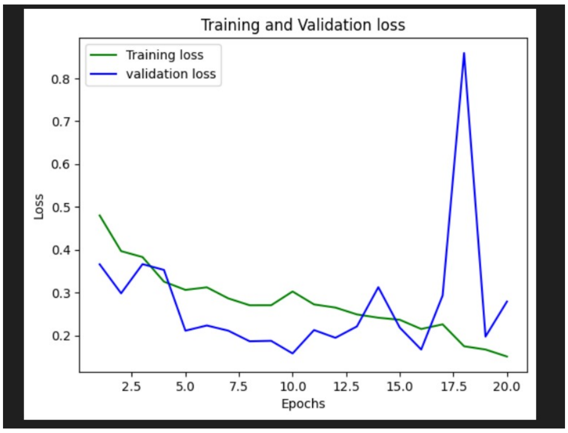

Figure 6.11: Loss

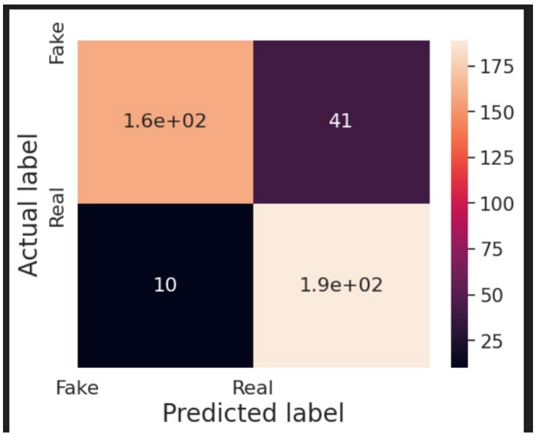

Figure 6.12: confusion matrix

Chapter 7

Conclusion and Future Scope

7.1 Conclusion

Our neural network-based approach successfully classifies videos as deepfake or real with a high level of confidence. Our method is capable of predicting the output by processing 1 second of video (10 frames per second) with a good accuracy. We implemented the model by using pre-trained ResNext CNN model to extract the frame level features and LSTM for temporal sequence processing to spot the changes between the t and t-1 frame. Our model can process the video in the frame sequence of 10,20,40,60,80,100.

This approach overcomes challenges faced by previous deepfake detection models, such as data oversampling issues, struggles with higher-resolution videos, and a lack of robustness. Our method empowers users to discern the authenticity of videos with improved accuracy and reliability.

7.2 Future Scope

There is always room for improvement in any developed system, especially when leveraging the latest trending technology with promising future prospects.

- **Upscaling to Browser Plugin/ Web Application:**This project can be scaled up from developing a web-based platform to a browser plugin for automatic deepfake detection. Even large applications like WhatsApp and Facebook can integrate this project into their applications for easy pre-detection of deepfakes before sending them to another user.This enhancement would provide users with more convenient and accessible deepfake detection capabilities.
- **Expanding Detection Capabilities:** While our current algorithm focuses on detecting face deepfakes, there is potential for enhancement to detect full-body deepfakes as well. This expansion would significantly improve the overall effectiveness and coverage of our deepfake detection system.

7.3 References

[1]Deepfake Detection: Current Challenges and Next Steps
Author:Siwei Lyu

[2]Towards Solving the Deepfake Problem: An Analysis on Improving Deepfake Detection using Dynamic Face Augmentation
Authors: Sowmen Das, Selim Seferbekov, Arup Datta, Md. Saiful Islam, Md. Ruhul Amin.

[3]Robust Face-Swap Detection Based on 3D Facial Shape Information
Authors: Weinan Guan, Wei Wang, Jing Dong, Bo Peng, Tieniu Tan

[4]Adversarially Robust DeepFake Media Detection Using Fused Convolutional Neural Network Predictions
Authors: Sohail Ahmed Khan, Dr. Alessandro Artusi, Dr. Hang Dai.

[5]Swapped face detection using deep learning and subjective assessment.
Authors:Xinyi Ding, Zohreh Raziei, Eric C. Larson, Eli V. Olinick, Paul Krueger, Michael Hahsler.

[6]Quick Overview of Face Swap Deep Fakes
Authors: Tomasz Walczyna and Zbigniew Piotrowski

[7]Digital Face Manipulation and Detection