Vigilens: Unmasking Deepfakes using Deep Learning

Trithi Amin, 211002

Sanskruti Bagul, 211005

Anusha Goyal, 211017

Department of Computer Science and Technology
Usha Mittal Institute of Technology
SNDT Women's University, Mumbai

April 16,2024

Overview

Abstract

- The growing computation power has made the deep learning algorithms so powerful that creating an indistinguishable human synthesized video popularly called as deepfakes has become very simple.
- Deepfake creation tools leave distinctive artefacts in the resulting deepfake videos which can be effectively captured by CNN.
- The algorithm uses CNN to extract the frame-level features and these features are further used to train the LSTM based RNN to classify whether the video is subject to any kind of manipulation or not.
- The primary aim of the project is to develop a vigilant system that can discern between authentic and manipulated media.

Introduction

- 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.

Problem Statement

- The primary objective of the deepfake detection system is to provide a reliable and scalable solution for identifying deepfake videos across various platforms and applications by training the model on diverse datasets containing both real and deepfake media samples.
- The system aims to learn discriminative patterns that distinguish between authentic and manipulated content.

Literature Survey

No.	Paper Name, Author(s), Year of Publication	Methodology and Technologies	Observations/ Findings and Remarks
1.	Adversarially Robust Deep-Fake Media Detection Using Fused CNN Predictions. Authors:Sohail Ahmed Khan, Dr.Alessandro, Dr.Hang Dai YOP: 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 proposed technique of the fusion model achieves 99 percent accuracy on lower quality Deep-FakeTIMIT dataset videos and 91.88 percent on higher quality DeepFake-TIMIT videos.

Literature Survey

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 YOP: 2018	The paper discusses various methodologies aimed at enhancing Deepfake detection accuracy like dynamic face augmentation, identifying dataset issues, face clustering, pre-processing guidelines.	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.

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. YOP: 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.	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 .

Literature Survey

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.	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.

Proposed System

Step 1: Data-set Gathering and Analysis

This step involves downloading the dataset relevant to deepfake videos, and preparing it for pre-processing.

Step 2: Module 1 Implementation

It 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

This step includes creating a new dataset that contains face-cropped videos for further model training.

Proposed System

Step 4: Module 2 Implementation

This step involves developing a data loader 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.

Workflow

- 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.

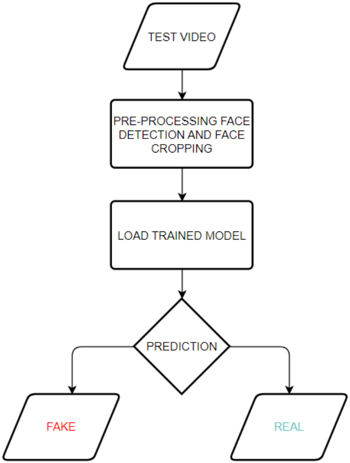
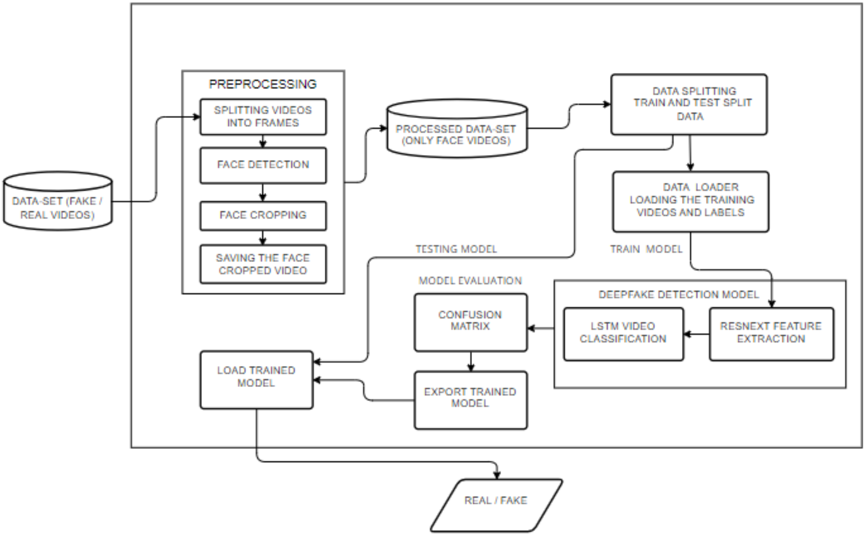


Figure: Testing Workflow

Architecture



Software and Hardware requirements

1.1 Hardware

- Intel Xeon E5 2637: 3.5 GHz
- RAM: 16 GB
- Hard disk: 100 GB
- Graphic card: NVIDIA GeForce GTX Titan (12 GB RAM)

1.2 Software

- Operating System: Windows 7+
- Programming language: Python 3.0
- Framework: Pytorch 1.4
- Libraries: OpenCV, Face-recognition
- Tool: Vs code, Jupyter

Dataset Gathering

- The primary objective was to acquire a dataset that encompasses diverse facial expressions, lighting conditions, and scenarios commonly encountered in real-world applications.
- Celeb DF v2 datasets are notable for their large size, containing a substantial number of videos featuring celebrities, and encompass a wide range of deepfake variations.
- FaceForensics is a video dataset that contains over 500,000 frames with faces from 1004 videos that can be used to research images or video forgeries. All videos are downloaded from YouTube and edited into short, continuous snippets with predominantly frontal faces.
- Celeb DF v2 datasets and Face forensic datasets both offer a rich variety of deepfake instances, including different manipulation techniques and degrees of realism, providing a robust foundation for training and evaluating our deepfake detection models.

Snippet of Dataset

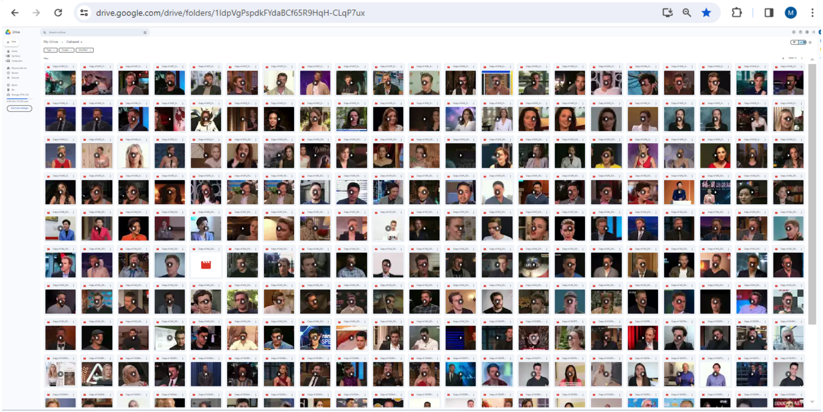


Figure: Snippet of Celeb VF2 Dataset

Preprocessing

- Using `glob`, all videos in the directory were imported into a Python list.
- `cv2.VideoCapture` read the videos and determined the mean number of frames, selecting 150 frames for uniformity in the new dataset.
- Videos were split into frames and cropped to the face location.
- Cropped face frames were written to a new video using `VideoWriter`.
- The new video was created with a resolution of 112 x 112 pixels at 30 frames per second in mp4 format.
- The first 150 frames were used to ensure proper use of LSTM for temporal sequence analysis.

Model Creation

- Combination of CNN and RNN
- Pre-trained ResNext CNN for feature extraction
- LSTM for video classification

ResNext:

- Used pre-trained ResNext for feature extraction
- Optimized for high performance on deeper neural networks
- Fine-tuning with extra layers and proper learning rate

LSTM for Sequence Processing:

- Input: 2048-dimensional feature vectors
- 1 LSTM layer with 2048 dimensions and 2048 hidden layers
- Leaky ReLU activation function
- Linear layer for correlation learning
- Adaptive average pooling layer for output size
- Batch size of 4 for training
- SoftMax layer for prediction confidence

Model Training

- **Train Test Split:** Dataset split into train and test sets with a 70-30 ratio.
- **Data Loader:** Loads videos and labels with a batch size of 4.
- **Training:** 20 epochs with a learning rate of 1e-5 (0.00001) and weight decay of 1e-3 (0.001) using Adam optimizer.
- **Adam Optimizer:** Enables adaptive learning rate.
- **Cross Entropy:** Loss function for classification problem.
- **Softmax Layer:** Final layer for probability interpretation (REAL or FAKE).
- **Confusion Matrix:** Summary of prediction results, evaluating model accuracy.

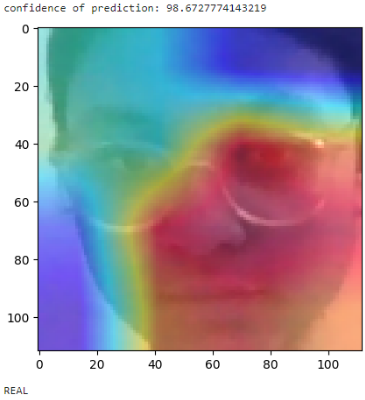


Figure: Real Video Output

- 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.

Prediction and Result

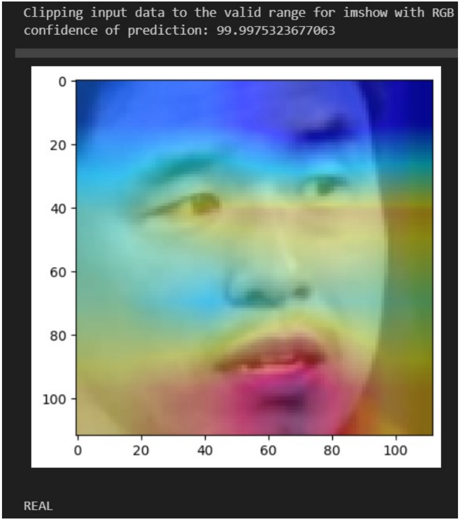


Figure: Real Video Output

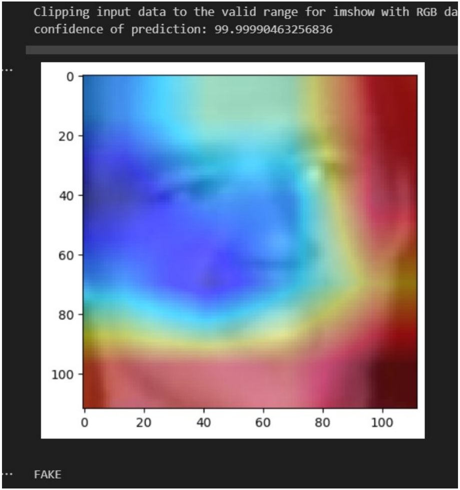


Figure: Fake Video Output

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.

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 (20 frames per second) with good accuracy.
- We implemented the model using a pre-trained ResNext CNN model to extract frame-level features and LSTM for temporal sequence processing to spot changes between the t and t-1 frame.
- This approach overcomes challenges faced by previous deepfake detection models, such as struggles with higher-resolution videos, data oversampling issues, and a lack of robustness.

References

- **Deepfake Detection: Current Challenges and Next Steps**
  Author: Siwei Lyu
- **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.
- **Robust Face-Swap Detection Based on 3D Facial Shape Information**
  Authors: Weinan Guan, Wei Wang, Jing Dong, Bo Peng, Tieniu Tan
- **Adversarially Robust DeepFake Media Detection Using Fused Convolutional Neural Network Predictions**
  Authors: Sohail Ahmed Khan, Dr. Alessandro Artusi, Dr. Hang Dai.
- **Swapped face detection using deep learning and subjective assessment**
  Authors: Xinyi Ding, Zohreh Raziei, Eric C. Larson, Eli V. Olinick, Paul Krueger, Michael Hahsler.

THANK YOU