## Deepfake Detection Project

Members:

Agam (102103151)

Chahat (102103160)

Arnav Sharma (102103162)

Umang Srivastava (102103165)

Mannat Kaur (102103176)

Shree Krishan Singla (102103178)



ABSTRACT

The emergence of deepfake videos through the use of Generative Adversarial Networks (GANs), which are capable of producing highly realistic photos and videos, has made it increasingly difficult to distinguish between real and fabricated digital images. To address this issue, a new deep learning model has been developed with the aim of efficiently differentiating between authentic and deepfake videos. The research utilized transfer learning techniques from computer vision, which involved utilizing previously trained features of neural networks used for image categorization to build a new detection model. As deep learning continues to evolve, both in the generation and detection of deepfakes, it is crucial to continually update detection techniques to remain effective. The results of the study are promising, with a detection accuracy of over 79%, and the potential for further advancement and development in this area.

INTRODUCTION

Over the past decade, the field of machine learning (ML) has experienced explosive growth, and its technology has been embraced by various industries for numerous practical applications. However, the continuous evolution of ML technology poses new data protection and security challenges. With the surge of online platforms that intake and broadcast videos, there is a pressing need to test the validity of these videos. Innovative approaches are necessary to tackle any confusion that may arise.

Deepfakes, which are videos or pictures that have been altered using artificial intelligence to appear different from their original state, were initially developed to create synthetic videos for the entertainment sector. However, with the advancement of deep learning, the technology behind deepfakes has been widely used to produce realistic doctored images and videos. The use of deepfakes has significant implications for privacy, anonymity, and mass communication. Consequently, big tech companies such as Facebook and Google have arranged competitions to identify and create datasets for deepfakes.

To address the challenges associated with deepfakes, technical experts have been developing new technologies. For instance, face detection using OpenCV is a popular technology for detecting faces in images and videos. Long short-term memory (LSTM) is another technology that can be used to analyse sequences of data and is particularly useful for modelling longer-term dependencies, such as facial expressions. Xception Net is another technology that is designed to learn essential features from face images.

Recycle-GAN is a technique that can be used to improve the accuracy of deepfake models. This technology allows the model to learn from its mistakes and generate more accurate results. Additionally, Recycle-GAN can be used to reduce the amount of data required to generate high-quality deepfakes.

Overall, the development of deepfake technology has had both positive and negative implications for society. While the technology has been used for entertainment and creative purposes, its malicious use has significant consequences for privacy, security, and trust. Therefore, researchers and technical experts are continuously working on developing new techniques to detect, prevent, and mitigate the harmful effects of deepfakes.

Bottom of Form

LITERARY SURVEY

In recent years, with the rise of advanced artificial intelligence, the methods for generating and detecting deepfakes have continued to evolve and become more sophisticated. The research community has been working tirelessly to improve deepfake detection algorithms and has published numerous findings. There is a growing struggle between those who use advanced machine learning to generate deepfakes and those who aim to identify deepfakes from real videos. As the consistency in the generation of deepfakes increases, the efficiency of the detection system needs to be enhanced accordingly. It has been suggested that what artificial intelligence has destroyed can also be restored by artificial intelligence.

Convolutional Neural Networks (CNNs) have been a major success in computer vision systems for supervised learning. Researchers have presented a new class of CNNs that render arithmetic operations feasible for filters between photos utilizing the latent vector by adding CNN models to GANs, generating cleverer forgeries. Deep Convolutional Generative Adversarial Networks (DCGANs) have also proven to be suitable for unsupervised study in the field of computer vision.

Initially, deepfakes were designed for the imitation of celebrities telling funny things. However, with the increase in accuracy, their applications have expanded to disrupting peace and prosperity by distributing fake news and generating community ruckus. The word “Deepfake” can have several meanings, but in this report, it refers to a video that includes a swapped face and is generated using a deep neural network. This is in contrast to "cheapfakes," which are fake videos produced with readily accessible software that has no learning part.

Researchers have theorized that deepfakes could be used to deliberately obtain other people’s knowledge by synthesizing faces or different parts of the body in videos. The authenticity of a video can be established by monitoring major shifts in eye blink patterns in deepfakes using a heuristic approach. The majority of video testing samples used for deepfake identification have a small number of faces closed with their heads, with an average intensity of 4.5 blinks per second, with each wink lasting approximately a quarter of a second. Failure to open eyes can also be a positive predictor of an in-depth analysis of deepfake videos.

Studying the most popular deepfake generation algorithms shows that they can produce fake faces with a specific size and resolution. To balance and suit the source face structure on the original images, an affine transformation and a blur feature need to be applied to the synthesized faces. An algorithm focused on finding inconsistency in the above applications can help in the detection of synthetically developed deepfakes.

The process of detecting GAN images produced using color indications often involves adding color discrepancy between the original images and the images created by GAN. However, this approach works for the entire video file and excludes the examination of different areas, as in the case of deepfakes.

METHODOLOGY

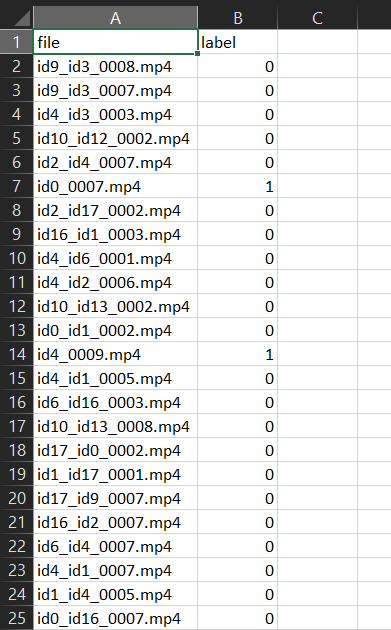
DATASETS:

We obtained our dataset, Celeb-DF, from the University at Buffalo, State University of New York, USA. The dataset contains both real and manipulated data, which we combined into a .csv file. The file consists of 954 rows and 2 columns, with headers 'file' and 'label'. The 'file' header contains the file name, and the 'label' column contains values of 0 and 1, where 0 represents fake data and 1 represents real data. We used this CSV file to train our model to differentiate between the two.

Here are some screenshots of the same:

Real Video

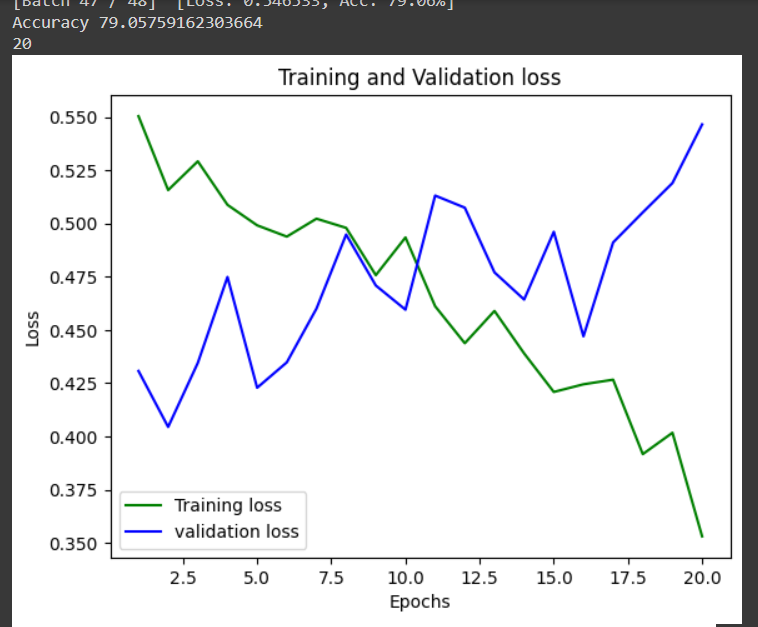
Fake Video

.csv file

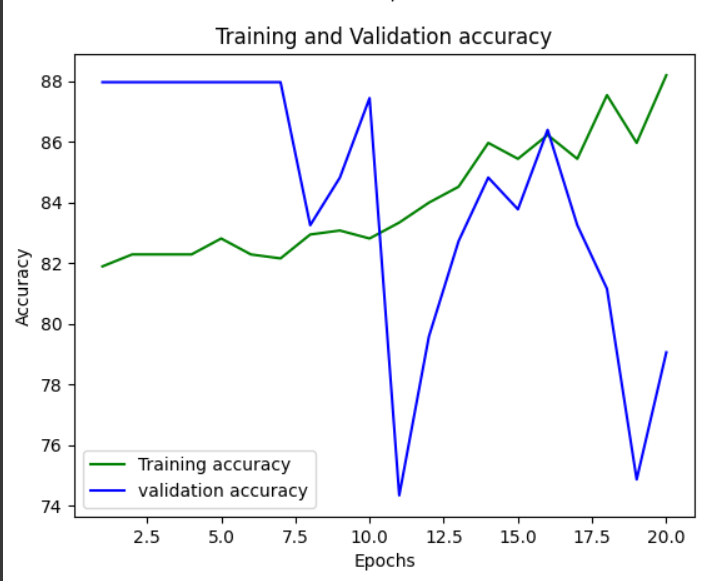
ALGORITHMS:

The algorithm used for this task involves OpenCV for video processing. Firstly, we counted the number of frames in each video and discarded any data that did not contain enough frames for our model to be trained effectively. We then extracted frames from the videos and transformed them into a format that would be compatible with our model. This involved converting the frames to a PIL image format, which gave us access to a number of functionalities. Next, we resized each frame to 112x112 and converted it to a tensor, which is necessary for faster image processing and to make it easier for the model's layers to extract features and make predictions. We then normalized the frames by reducing their mean to 0 and their variance to 1, as high variance can result in certain weights being given more importance than others. This transformed data was then provided to the data loaders of PyTorch to help with batch processing. Our pre-trained model, ResNet, from the torchvision library, was used to extract features and train the model. We removed the last two layers of the model, which were intended for object detection, and added an LSTM layer and a layer that used LeakyReLU to train the model. We used LeakyReLU to overcome the problem of dying ReLU. We also used the dropout layer to avoid overfitting, which can negatively impact the model's accuracy and performance. AdaptiveAvgPool2d was used to reduce the dimensions to a single value so that it could be sent to LSTM. We trained our epoch, calculated the losses and accuracies, and backpropagated to reduce losses and increase accuracy. We then constructed a confusion matrix containing the true positives, false positives, false negatives, and true negatives.

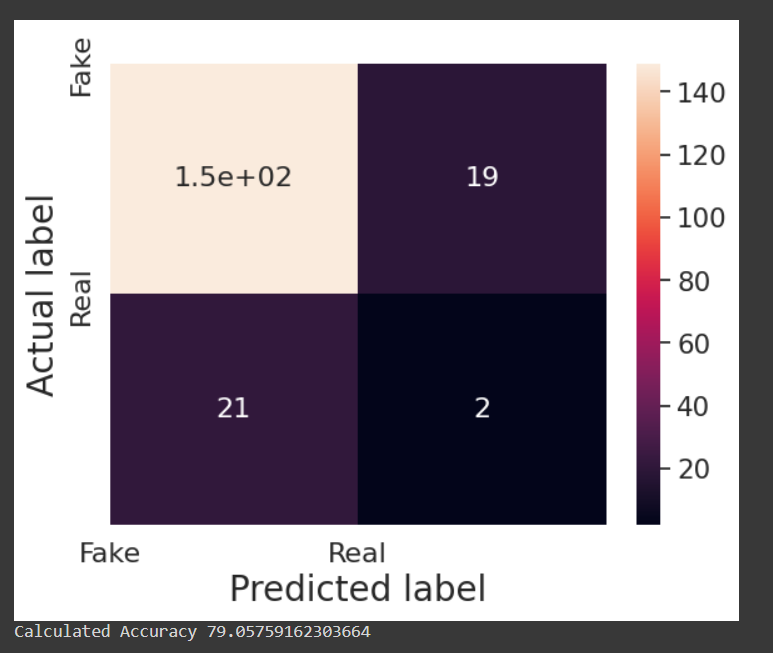
RESULTS



The plot indicates that the training loss is decreasing while the validation loss is increasing, which is not a good sign. This could be due to overfitting of the model to the training data, or it may not be able to generalize well to new data. To improve the model's performance, it may be necessary to collect more data.



The same trend is observed in the accuracy as well. To address this issue, it may be necessary to collect more data and adjust the model's hyperparameters.



Here is the confusion matrix, which shows that there are 150 true positives, 19 false positives, 21 false negatives, and 2 true negatives. The calculated accuracy is 79.05%.

FUTURE SCOPE

The primary obstacle encountered in this study was the insufficient computational power. Since the research involves video datasets, a significant amount of RAM is necessary to store the data undergoing analysis. Several techniques were employed, such as optimizing the model to run minimal analyses and reducing video sizes and factors, in an attempt to directly train the model on the video dataset. However, the model ran out of memory in each case.

As a result of this limitation, the model had to be trained on images of frames captured from videos. This shortfall in computation power resulted in decreased efficiency in detecting deepfake videos for a model that could learn video categorization directly. Nonetheless, this research presents a tremendous opportunity for future work. The dataset utilized in this research represents a small fraction of data from a challenge organized to develop a deepfake detection model. The complete dataset is a vast 950GB that cannot be trained on a local machine or even on Google Colab.

After attempting various methods, it can be inferred that detecting deepfake videos presents several challenges. Additionally, training with limited videos is inadequate for developing a comprehensive model. With greater computational power and memory resources, a more advanced model can be developed that can take videos as input.