Deepfake Video Detection

# Shankar Bhawani Dayal; Brett van Niekerk University of KwaZulu-Natal; South Africa sbdayal,@gmail.com vanniekerkb@ukzn.ac.za

## 2021

This paper aims to determine if deepfake models which are pre-trained on older datasets are still able to accurately detect whether a video is real or a deepfake from a newer dataset

# Scholarcy Highlights

* Deepfake media, such as videos and images, are the greatest form of crime created from artificial intelligence (Smith, 2020)
* This project has successfully tested a popular and well respected paper by Rössler et al (2019) and found the accuracy is potentially not as high as it is claimed to be, the believed reason for the accuracy being so low is that the deepfakes in the Faceforensics++ dataset are firstly, not big enough in terms of the amount of videos present in the dataset, secondly, does not accurately represent each scenario in which a deepfake video may occur
* The Faceforensics++ dataset chose videos where the targets face is front facing, it would be expected for any model trained on only front facing deepfakes to not be able to accurately or effectively detect deepfake videos when the target is not front facing
* This project shows the evident need for constant testing of previous work to determine if and when they are no longer sufficiently adequate at detecting deepfake material, which indicates that the fight against deepfakes may be an unending task
* The Faceforensics++ dataset is approximately 3.5TB in size, and it was not downloaded due to lack of storage space. This project can be trained on the Deepfake Detection Challenge (DFDC) dataset and tested on the Faceforensics++ dataset to determine if the accuracy achieved is greater than trained on the Faceforensics++ dataset and tested on the DFDC dataset, as this project accomplished
* Güera and Delp (2018) tested their network against 20, 40 and 80 frames from each video. They state that they have 96.7%, 97.1% and 97.1% accuracy respectively when they tested it
* The effectiveness of a deepfake detection model seems to be highly dependent on the quality of the dataset, this might suggest that deepfake detection models may always be proved to be ineffective once a newer deepfake creation method is used, since methods which utilize the Generative Adversarial Networks (GANs) architecture may utilize a deepfake detection model as the discriminator to create the new generation of deepfake material

# Scholarcy Summary

## Introduction

Deepfake media, such as videos and images, are the greatest form of crime created from artificial intelligence (Smith, 2020).

The idea of deepfakes was first proposed on Reddit by a user called “deepfakes” (Fink and Diamond, 2020; Sample, 2020), who uploaded videos of his work which happened to be deepfake pornographic material of famous Hollywood actresses.

GANs utilize two neural networks working against each other, where the “generator” network creates material trying to fool the second network, and the second network which is the “discriminator” attempts to spot the real or fake instances.

This is the adversarial component of the network.

The generator and discriminator learn from each other, since the generator's output is linked directly to the discriminators input, and when backpropagation is occurring, the discriminators prediction/output is used as a signal for the generator to update its weights (Google Developers, 2019)

## Methods

The computer used for this project has: Operating System: Windows 10Processor: Intel(R) Core (TM) i5-7300HQ CPU @ 2.5GHzGraphics Card: Nvidia GeForce GTX 1050Installed RAM: 8GBSystem Type: 64bit Operating System, x64-based processor3.1 Creating the datasetThe first step is creating the sub-dataset from the DFDC dataset.

System Type: 64bit Operating System, x64-based processor.

Two folders are created for each segmented dataset, i.e there are six total folders.

A python program is created to open the metadata.json file that is in each segmented folder, create a list of the fake videos present in the folder by selecting each video where label equals “FAKE” in the metadata.json file.

This list of fake videos has the name of each video present in the segmented folder, in a for-loop for each video in the list: Add “\\” and the name of the video to the end of the file path of the segmented folder, this gives you the file path of the video

## Results

Results and discussionAfchar et al (2018) was not able to be tested, to no fault of the code or the model, this is an error with Keras, as such, there is no way we can validate or verify the effectiveness of the model on a portion of the DFDC dataset.Rössler et al (2019) with pretrained weights achieved 13.1578947368% accuracy for detecting deep fakes and 89.4736842105% for detecting real videos, with an overall accuracy of 51.3157894736%.Classification accuracy for deepfake videos: Classification accuracy for realvideos: Pretrained XceptionNet tested on DFDC datasetXceptionNet results from Rössler et al (2019) UnknownAverage classification accuracy: 82.01% accuracy on raw uncropped video footage0.0% fake aarsmohwrt.mp40.0% fake aayrffkzxn.mp40.0% fake abxtkdjyru.mp40.0% fake afnxnrrqsj.mp40.0% fake aheocfkxjx.mp.

Afchar et al (2018) was not able to be tested, to no fault of the code or the model, this is an error with Keras, as such, there is no way we can validate or verify the effectiveness of the model on a portion of the DFDC dataset.

Rössler et al (2019) with pretrained weights achieved 13.1578947368% accuracy for detecting deep fakes and 89.4736842105% for detecting real videos, with an overall accuracy of 51.3157894736%.

Classification accuracy for deepfake videos: Classification accuracy for realvideos: Pretrained XceptionNet tested on DFDC dataset.

XceptionNet results from Rössler et al (2019) Unknown Average classification accuracy: 82.01% accuracy on raw uncropped video footage

## Conclusion

This project has successfully tested a popular and well respected paper by Rössler et al (2019) and found the accuracy is potentially not as high as it is claimed to be, the believed reason for the accuracy being so low is that the deepfakes in the Faceforensics++ dataset are firstly, not big enough in terms of the amount of videos present in the dataset, secondly, does not accurately represent each scenario in which a deepfake video may occur.

The methods used to create the deepfakes in the Faceforensics++ dataset are seen in other papers such as in Afchar et al (2018), the premise is that they will have deepfakes of the same quality which would not adequately train them to detect deepfakes that are created with methods seen today.

This project shows the evident need for constant testing of previous work to determine if and when they are no longer sufficiently adequate at detecting deepfake material, which indicates that the fight against deepfakes may be an unending task