Evaluating Deepfake Video Detection Algorithms: A Transition From Metrics to Adversarial Attacks

Omkar Sarde
Rochester Institute of Technology
1 Lomb Memorial Dr, Rochester, NY 14623
os4802@rit.edu

Abstract

Falsified imagery in pictures and videos created through deep learning techniques is collectively termed as deepfakes. An increase in computing power coupled with multiple open-source implementations of deepfake imagery generators have resulted in an extensive proliferation of deepfake imagery. Ease of creation, utilization, and difficulty of identification make deepfake imagery, especially deepfake videos, extremely potent tools in malicious actors' arsenal. Multiple deepfake video detection algorithms exist to counter the threat of deepfake videos. These detectors utilize various variants of the same underlying methodologies to detect deepfake videos and promise to be the best among their peers. Such claims coupled and reported metrics have led to ambiguity regarding the efficacy of deepfake detectors.

This piece of literature aims to mitigate the ambiguity in reported metrics of two prominent deepfake detector families by creating a benchmarking methodology to evaluate them. With the inclusion of robustness criteria against adversarial attacks and constraints on hardware capabilities, we attempt to make the benchmark more generalizable. Finally, we conclude this piece by contrasting the benchmark results with detectors' reported metrics under consideration.

1. Introduction

Deepfakes are falsified imagery, especially videos generated through perturbations to original imagery by utilization of standalone or combination of Deep Learning (DL) based techniques. Deployment of applications to generate deepfakes like 'deepnude' [44] have proven their potential to disrupt societal fabric through gross violation of privacy and propagation of misinformation. Significant growth has been observed in the figures of deepfake generators and resultant imagery, with recent reports depicting a 330% growth of online deepfake videos since July 2019 [1]. Concerning growth in generation and prevalence of the existence of deepfakes has provided an impetus to research on combating deepfakes. Nguyen *et al*. [33] estimate a 138% annual growth on deepfake related research since 2019 and present a comprehensive survey on existing deepfake creation and detection techniques.

Most deepfake detection techniques rely on traditional metrics to evaluate their efficacy on particular datasets of their selection. Lack of uniform benchmarking methodology coupled with a deficit of datasets has resulted in varying estimates of the efficiency of deepfake detection tools. This project attempts to solve this problem through the following contributions.

- We evaluate primary classification metrics of predominant deepfake detection techniques that employ Convolutional Long Short Term Memory (CLSTM) Residual Networks [43, 19] and Eye Blinking [28] on latest iterations of two datasets [29, 11].
- We evaluate the robustness of the chosen models against adversarial attacks under the white box and black box settings [32, 15].
- We propose a benchmarking methodology for evaluation of the deepfake detection techniques based on a hybrid combination of chosen datasets and adversarial attacks and compare it with the original metrics provided by the models under consideration.

The following sections cover the work done in deepfake video detection, the methodologies used, the implementations and evaluation measures along with the legal and ethical issues. Section 2 discusses the work done in the deepfake video detection field over the years and the prerequisites for the study. Section 3 proposes a benchmark methodology to test the two families of deepfake video detection. Section 4 pertains to the evaluation metrics that form the

crux of the proposed methodology for evaluation of deepfake video detectors. Section 5 emphasizes on the implementation architecture used to evaluate the baseline deepfake video detectors. Section 6 summarizes the results obtained from restrictions for evalution metrics using the proposed methodology. Section 7 discusses the possible reasons for deviations from the reported metrics from our study and the baseline methods. Section 8 discusses the associated legal and ethical issues. Finally the paper is concluded by final remarks and possible future direction of deepfake video detector development in section 9.

2. Related Works

2.1. Deepfake Imagery Generation

The primary object of deepfake creation can be summarized as transferring localized features of an input source to a target source. In the case of Imagery, particularly videos, the transfer features consist of facial expressions and important facial landmarks in the form of eyes, nose, lips, etcetera. The meteoric rise in deepfake related research [1] has led to the development of sophisticated algorithms for deepfake imagery generation. Recent algorithms devised for deepfake generation employ complex techniques to generate photo-realistic artificial face videos [4, 8, 10, 21, 22, 25, 27, 31, 41, 42, 45, 46].

Majority of deepfake imagery generation algorithms are based on variations of the base work of neural image style transfer technique [30]. The base framework utilizes an Unsupervised image-to-image Translation (UNIT) framework based on generative adversarial networks (GANs) and variational autoencoders(VAEs) to generate a translation of features from input to target imagery. A detailed description of this deepfake autoencoder (DFAE) technique is explained in section 2.1.1. The resultant deepfake imagery developed by this method demonstrated excellent quality yet had two limitations. The first limitation was in the form of the translation model is unimodal due to the Gaussian latent space assumption and the second limitation was in the form of unstable training due to the saddle point searching problem. Even with it its limitations, acceptable quality of generated fakes, and the ability to scale of this technique have led to several its independent open-source implementations in the form of DeepFaceLab [16], DFaker [17], Faceswap [13], Faceswap-GAN [12] and FakeApp [14].

Another noteworthy category of deepfake imagery generation algorithms employ algorithms incorporating GANs. Neural Talking Heads (NTH) model [51], Face Swapping GAN [34] and StyleGAN [23] are examples representing this category. B. Dolhansky *et al*. [11] provides an exhaustive comparison between the two major categories of deepfake video generation. Their experiments found models based on GAN-like methods to work well in restricted settings necessitating proper lighting conditions but these models were not as effective as the DFAE based models. The DFAE based models outperformed GAN-like models through the generation of higher quality deepfakes over a wider range of settings and were relatively easier to employ. These advantages of DFAE based methods over GAN-like methods are reflected through the growing popularity of DFAE based methods and their easily available scalable implementations.

2.1.1 Deepfake Autoencoder based Imagery Generation Pipeline

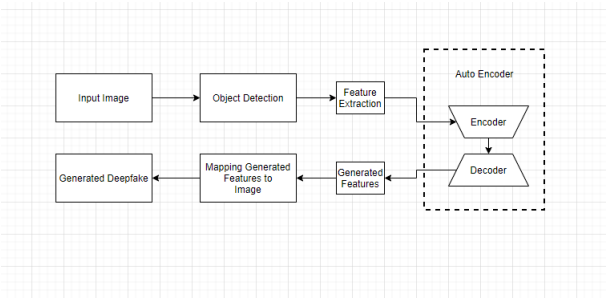

Figure 1. Basic Deepfake Imagery Generation Pipeline

The main objective of a deepfake imagery generation pipeline is the translation of selected features from an input source to a target source. The chosen features are majorly facial landmarks in the form of eyes, nose, lips, etcetera, and are mapped to target sources to generate realistic synthetic imagery. The most popular methods to implement such a pipeline are based on a neural style transfer technique utilizing variational autoencoders (VAEs) [30, 29, 11]. These methods can be summarized under the umbrella term of deepfake autoencoder (DFAE) based methods due to their reliance on VAEs.

Variety of variations of the DFAE based models for deepfake imagery generation exist in the form of open-source implementations [16, 17, 12, 13, 14]. The underlying mechanism for these models are similar and following a basic pipeline. A basic deepfake maker utilizing DFAEs is shown in Fig. 1. The input to the pipeline is a video consisting of faces of the target individuals that are to be translated to output fake imagery. The input is then passed through object detection algorithms to identify features of interest. Important features in the form of facial landmarks like eyes, nose, lips are then extracted from the output of the object detector. The extracted features are then aligned and standardized to a standard congfiguration [24]. Standardized and aligned extracted features are then passed to an autoencoder [36] in the final stage of the pipeline to create synthetic imagery with similar facial features and expressions as the input source and mapped on a donor source.

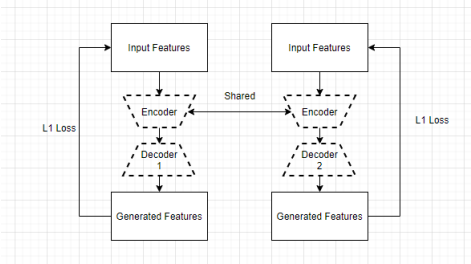

Figure 2. AutoEncoder Architecture

The autoencoder stage comprises usually of two convolutional neural networks (CNNs) for the encoder and decoder components respectively. E converts the standardized aligned features provided as input to the autoencoder to a vector. A single encoder is used irrespective of several subject identities to guarantee identity-independent attributes in the form of facial expressions. Multiple decoders are utilized to generate faces (output imagery) of each corresponding subject identity (source imagery). The encoder-decoder is trained consecutively consuming multiple faces from the input datasets in an unsupervised manner. The decoder generated faces are then translated to the donor imagery using a variety of methods such as warping the configuration of the output donor face with the decoder generated face.

2.2. Deepfake Video Datasets

The rise in the amount of deepfake video imagery has garnered research into devising accurate detection tools to identify deepfakes. The prime component required in development of these detection tools is large, high quality datasets that contain deepfake imagery over a large variety of subjects in various setting for training of the detectors. This has led to development of various deepfake video imagery datasets in recent years.

Li *et al*. [29] were first to categorize the deepfake datasets into two generations based upon the criteria of size and imagery quality. B. Dolhansky extend the philosophy of this dataset generation categorization *et al*. [11] by addition of a third generation based upon the consent of subjects used to create the imagery. According to their categorization deepfake video imagery datasets can be classified into three generations. The first generation of deepfake video datasets consist of the DF-TIMIT [26], UADFV [50] and FaceForensics++ DF [38]. The first generation datasets are characterized by smaller quantity and lower quality of videos compared to the later generations. The second generation of the deepfake datasets comprised of Celeb-DF [29] and the Google DeepFake Detection Dataset [5]. The second generation of deepfake datasets mitigated the flaws of the first generation datasets through increasing number of videos and incorporating enhancement to the DFAE architecture used for synthesizing deepfakes. The third generation of the deepfake datasets presently consists of the DeepFake Detection Challenge (DFDC) dataset. It is currently the largest dataset for deepfake video imagery and is characterized by consent provided by actors who contributed in creation of the dataset.

2.2.1 Celeb-df

Y. Li *et al*. [29], created the Celeb-DF dataset, a second generation deepfake dataset, with an objective of mitigating the flaws of lower resolution quality and a small amount of videos present in the first generation of deepfake datasets. To achieve its objective the Celeb-DF dataset enhances the basic deepfake autoencoder (DFAE) based method for generation of deepfakes mentioned in section 2.1.1. by incorporating techniques to enhance resolution quality. The first refinement to DFAE is done through increasing number of layers in the encoder decoder architecture to increase resolution. Color mismatch between donor and target faces are mitigated through training data augmentation and utilization of a color transfer algorithm [37]. Incorporation of proper transfer of features from target to donor imagery is achieved through smoothing in the form of utilization of advanced face masks. Finally, temporal flickering is reduced through usage of temporal correlations among facial features. The Celeb-DF dataset consists of 5,639 unique fake videos and 6,229 total videos using a total of 59 subjects.

2.2.2 DeepFake Detection Challenge Dataset

B. Dolhansky *et al*. [11] created the DeepFake Detection Challenge (DFDC) Dataset as an improvement to the second generation of deepfake datasets. DFDC pays particular emphasis on impact of deepfakes on subjects in source imagery used to create the dataset through being the first dataset to gain consent from contributing subjects. This was achieved through commissioning a large amount of videos from various individuals who provided explicit consent for the creation of the dataset. The provided consent coupled with the large dataset size make DFDC one of the few datasets in the third generation of deepfake datasets.

Although the exact information of the deepfake imagery generation models used for DFDC creation was not disclosed, the underlying methodologies were provided. DFDC employs both the major classes of deepfake imagery generation, namely the DFAE mentioned in section 2.1.1 and GAN-like methods of Neural Talking Heads (NTH) model [51], Face Swapping GAN [34] and StyleGAN [23]. The exact distribution of videos generated using particular methods was not provided yet it was stated that most of the videos were created through utilization of the DFAE due to inherent qualitative advantages of DFAEs over GAN-like methods. The DFDC consists of 104,500 unique fake

videos and 128,154 total videos using a total of 960 agreeing subjects making it the first and presently the only deepfake dataset with over a 100,000+ videos.

2.3. Deepfake Video Detection

As research on deepfake detection is on rise [33] multiple new models have been created for deepfake imagery detection in recent years. Majority of these deepfake imagery detection tools focus on identification of deepfake images rather than deepfake videos [33]. Video data comprise of temporal characteristics along with image characteristics reducing the efficacy of deepfake image detectors. Deepfake video imagery detectors utilize this temporal characteristic in standalone or in combination of other relevant image characteristics to identify deepfake videos.

Deepfake video detection models can be classified into two major classes based on their pipelines and features of interest. The first classes of deepfake video detectors use temporal features across video frames in a standalone manner. The base hypothesis of this subgroup of model relies on the fact that deepfake video generation pipelines create deepfakes on a frame-by-frame basis as mentioned in section 2.1.1. This makes identification of inconsistencies across frames in the temporal dimension a key element in discovery of deepfake videos. The models of this category are mainly formulated through utilization of a Convolutional Neural Network (CNN) for consideration of the visual characteristics and a Recurrent Neural Network (RNN), majorly in the form of Long Short Term Memory (LSTM) network to account for the temporal dimension.

The other class of deepfake video detectors utilizes knowledge of human anatomy to identify deepfake videos. These deepfake video detectors focus on facial artefacts, majorly by concentrating on the human eye region. This class of deepfake video detectors function by distinguishing fake facial features from real ones through employing domain knowledge of human anatomy, particularly eye blinking rate. Such models employ a similar framework as those of the deepfake video detectors using temporal features yet their focus on the human facial artefacts of eyes distinguishes them.

2.3.1 Deepfake Detection using Temporal Characteristics

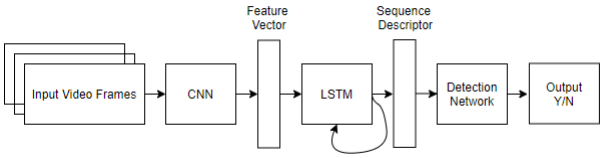

Figure 3. Basic CNN+LSTM architecture for deepfake detection

Tariq *et al*. [43] created a Convolutional Long Short Term Memory (LSTM) based Residual Network (CLRnet) model to detect deepfakes from [39] dataset with model generalizability as the focal area. Most deepfakes contain irregularities in the form of variation in the spectral qualities like brightness and contrast and variation in facial features of input at locations like eyebrows, eyes, and lips in between consecutive frames. CLRnet uses a convolutional LSTM block to capture these irregularities and uses them as a key classification feature. This convolutional LSTM block is then attached to a Residual Network to avoid the vanishing gradient problem. They validated their model on various datasets using direct and transfer learning approaches. The reported accuracy and metrics found CLRnet coupled with transfer learning achieved the highest accuracy amongst state-of-the-art models

Guera *et al*. [19] use a very similar approach to [43] without utilizing transfer learning characteristics. Surprisingly, even without transfer learning, they claim to report similar accuracy and performance metrics to [43] necessitating further evaluation of claims of both [43, 19] on the same datasets.

2.3.2 Deepfake Video Detection Using Facial Artefacts of Eye

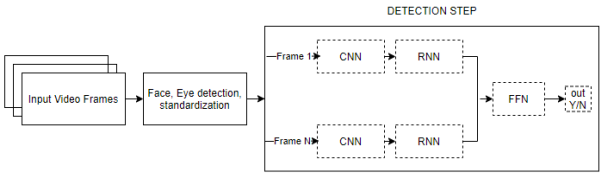

Figure 4. Basic deepfake detection using Facial Artefacts

The other major class of deepfake detection models utilizes a lack of blinking of human eyes as an essential classification feature. Most of these models are derivatives of the work of Li *et al*. [28]. The base model utilizes a physiological signal in the form of the blinking of human eyes to detect deepfake videos. The base model utilizes a Long-term recurrent convolutional network (LRCN) focused on the eye region to measure the blinking of human eyes.The fixation of the LRCN on the eye region is the key differentiator between Eye blinking based models and Convolutional Long Short Term Memory (LSTM) based models. Both of these classes attempt to leverage the sequential information in space and time domains encompassed between consecutive frames and report similar metrics of accuracy and performance.

2.4. Adversarial Attacks

The efficiency of deepfake video detection tools depends upon both, quality of the training data and the test data, that is the videos used to determine their accuracy. This dependency on data makes deepfake video detection tools extremely susceptible to adversarial attacks. Adversarial attacks are intentional perturbations to input provided to a machine learning model with an intent to cause misclassification by the model. The intentionally perturbed inputs are known as Adversarial Examples (AEs). Adversarial attacks are categorized in two categories of white-box attacks and black-box attacks. In the white-box attack setting the adversary has unhindered access to the model under attack. In the black-box attack setting the adversary only has API access to the model under attack. Black box attacks can be further classified into two major subsections of transfer attacks and optimization attacks.

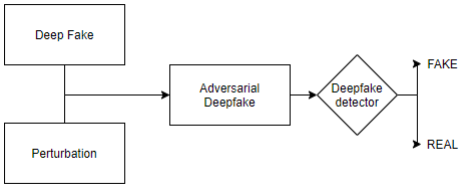

Figure 5. Basic Adversarial Attack Architecture

Deep fake video detector model vulnerability to AEs makes robustness and defense against AEs an important metric to judge deepfake detection techniques [6]. Adversaries may have complete access to the target model, white-box attacks, or only API access through model queries to the target model, black-box attacks, to generate AEs. Recent research by Gandhi*et al*. [15] used the Fast Gradient Sign Method (FGSM) [18] and Carlini and Wagner *L2* Norm attack (CWL2) [7] to stage both white boxes and transfer learning-based black-box attacks to demonstrate deepfake detection model vulnerabilities towards AEs.

Many implementation variations exist for FGSM method [18]. The core idea of this adversarial attack is based on exploitation of gradient of the training loss with respect to the input sample. The gradient loss is given by:

$$Gradientloss = \Delta_{\mathrm{x}} J(x, y, \theta) \tag{1}$$

and the adversarial example is given through the following equation:

$$X_{\mathrm{adv}} = x + sign(\Delta_{\mathrm{x}} J(x, y, \theta)) \tag{2}$$

wherein X_{adv} is the generated adversarial sample, x is the input sample and y the prediction of the model.

Carlini and Wagner *L2* Norm attack (CWL2) [7] is based on minimization of *L2* norm of the perturbation summarized as:

$$min_{\mathrm{x}}, ||x' - x||_2{}^2 \tag{3}$$

and the adversarial sample is given through the following equation:

$$w* = arg_{\mathrm{w}} min(||x' - x||_2{}^2 + cf(x')) \tag{4}$$

$$X_{\mathrm{adv}} = \frac{1}{2}(tanh(w*) + 1) \tag{5}$$

They also suggest defenses in the form of utilization of Lipschitz regularization [49] and Deep Image Prior [47] to negate the effect of perturbations to input images.

3. Proposed Benchmarking Methodology

Extant deepfake detection techniques utilizing Deep Learning (DL) frameworks can be classified into classes using Convolution based Long Short Term Memory (CLSTM) Networks [43, 19] and Eye Blinking [28]. These techniques estimate their efficacy on datasets of their choice, leading to variance in their reported metrics. Such datasets utilize automatic encoders and Generative Adversarial Networks (GANs) based methods [29, 11] to generate deepfakes. The efficacy metrics chosen by the models thus can be inferred to be dataset variant and require further investigation. Also, recent research indicates deepfake detection models to depict vulnerability to adversarial attacks under both black box and white box settings [15] necessitating robustness tests under such attacks to be an important performance metric to gauge model efficacy.

To mitigate this variance we propose using a combination of two of the largest deepfake datasets [29, 11] and sequential adversarial attacks from [15]. This would help generalize results among different models based on a combined dataset and ensure model robustness to Adversarial Attacks.

The proposed combined dataset would consist of 500 randomly chosen videos from each, the second generation dataset Celeb-DF, and the third generation dataset DFDC. To make the benchmark commodity hardware usable we restrict the number of videos to 500 per dataset and employ a single mid-range graphics processing unit. Random selection of videos would mitigate any induction of inadvertent user selection bias. The models under test would be created using the programming language of *Python* using the *PyTorch* framework to make the bench open-source framework compatible. Further, after model creation and training, we would test the models against adversarial examples (AEs) under both black-box and white-box settings. White-box AEs generated for one class of deepfake detection models would be employed as black-box AEs for the other class to minimize the induction of bias.

As multiple variations of metrics are possible to estimate the primary efficacy of the aforementioned deepfake video

detector model families and their robustness against adversarial attacks we use the evaluation metrics discussed in detail in Section 4 Evaluation Metrics.

Through randomly selecting limited videos spanning across two generations of deepfake datasets, restricting hardware requirements and minimizing induction of user selection bias at every step of the benchmark we aim to create a platform that can be utilized to benchmark the efficacy of deepfake video detectors in a transparent manner to aid deepfake detection research.

4. Evaluation Metrics

The focus of our study is to understand the variation in reported metrics of the two main families of deepfake video detectors, namely the deepfake video detectors leveraging only temporal characteristics discussed in Section 2.3.1 and the deepfake video detectors using facial artifacts of the eye discussed in Section 2.3.2. Multiple variations to the baseline detectors discussed above exisit [43, 28]. Also, even the implementation of the baseline methods is not provided in the base methods [19, 28]. The baseline methods do not share the same datasets and the models developed are tested on self-generated data, further increasing discrepancy. Furthermore, the reported models were run for the author's chosen time periods adding more ambiguity.

To account for ambiguity we have developed the following Evaluation Metric:

- We develop the base models as-is from the architecture reported in the baseline methods using open source architecture using the same set of hyper-parameters.
- To gauge the effect of base Convolutional Neural Network (CNN) used we train both the baseline detectors using VGG [40] and Resnet [20], particularly, using the VGG16 and Resnet50 models and using the same attached Long Short Term Memory architecture models for both of them.
- We train the base models using a dataset comprising of equal amounts of samples from two generations of deepfake datasets mentioned in Section 2.2 using commodity hardware based on 100, 500, 1000, 2000, and 3000 samples.
- To account for data distribution shift we draft three test sets consisting of only second-generation deepfake dataset Celebdf [29] discussed in Section 2.2.1, only third-generation deepfake dataset Facebook DFDC [11] discussed in Section 2.2.2 and a final test set made of equal samples comprising both the generations.
- Finally, we test the models against both white-box and black-box adversarial samples generated with FGSM [18] and Carlini Wagner L2 [7] methods discussed in Section 2.4.

By using this evaluation metric, we try to account for every possible ambiguity reflected through *'user selection bias'*. It is to be noted that we do not test the models for maximum possible accuracy due to present hardware limitations due to criteria of usage of commodity hardware. By inclusion of these criteria, we make sure that undue advantages of training hardware are mitigated.

5. Implementation Architecture

To gauge if discrepancies exist in the baseline methods mentioned in Section 2.3 either due to choice of architecture, size, and quality of dataset used to train and evaluate the model or due to a combination of both, commodity hardware was used to implement both the baseline methods.

As code implementations of both the baseline methods was not presented, we implemented them using the programming language of *python* using the *pytorch* framework. Both the models were created using two Convolutional Neural Network(CNN) of VGG16 and Resnet50, open-source available implementations of [40, 20]. The Long Short Term Memory used to process output from the CNNs was the same for both methods. The four models (one using VGG, one using Resnet for both baseline methods) were trained for 20 epochs each. The video data used as input was fixed as 20 frames for input length to avoid discrepancies. To process the input, we used *OpenCV* framework to generate frames from the video. To develop features of eye movements necessary for the 'Deepfake Video Detection Using Facial Artefacts of Eye' model, we use the *Haar Cascade* [48] implementation using the *OpenCV* framework. Finally, both the models were trained and evaluated for dataset conditions mentioned in Section 4 using an Nvidia 1660ti GPU [9] and AMD Ryzen 7 4800H CPU [2].

For a generation of adversarial samples we generate white-box adversarial using both Fast Gradient Sign Method (FGSM) [18] and Carlini Wagner L2 Norm Attack Method (CWL2) [7] for both the VGG16 and Resnet50 architectures. White-box adversarial for one architecture were used as black-box samples for the other architecture.

6. Results

Samples	Model	Test Set	Accuracy
100	Base1 VGG16	2nd Gen	79%
100	Base1 VGG16	3rd Gen	77%
100	Base1 VGG16	Combined	76%
100	Base1 Resnet50	2nd Gen	73%
100	Base1 Resnet50	3rd Gen	71%
100	Base1 Resnet50	Combined	70%
100	Base2 VGG16	2nd Gen	75%
100	Base2 VGG16	3rd Gen	72%
100	Base2 VGG16	Combined	72%
100	Base2 Resnet50	2nd Gen	69%
100	Base2 Resnet50	3rd Gen	67%
100	Base2 Resnet50	Combined	65%

Table 1. Results without Adversarial Attack 100 samples

Samples	Model	Test Set	Accuracy
500	Base1 VGG16	2nd Gen	81%
500	Base1 VGG16	3rd Gen	78%
500	Base1 VGG16	Combined	77%
500	Base1 Resnet50	2nd Gen	75%
500	Base1 Resnet50	3rd Gen	73%
500	Base1 Resnet50	Combined	71%
500	Base2 VGG16	2nd Gen	79%
500	Base2 VGG16	3rd Gen	74%
500	Base2 VGG16	Combined	72%
500	Base2 Resnet50	2nd Gen	73%
500	Base2 Resnet50	3rd Gen	70%
500	Base2 Resnet50	Combined	69%

Table 2. Results without Adversarial Attack 500 samples

Samples	Model	Test Set	Accuracy
1000	Base1 VGG16	2nd Gen	85%
1000	Base1 VGG16	3rd Gen	82%
1000	Base1 VGG16	Combined	83%
1000	Base1 Resnet50	2nd Gen	79%
1000	Base1 Resnet50	3rd Gen	77%
1000	Base1 Resnet50	Combined	76%
1000	Base2 VGG16	2nd Gen	82%
1000	Base2 VGG16	3rd Gen	78%
1000	Base2 VGG16	Combined	75%
1000	Base2 Resnet50	2nd Gen	77%
1000	Base2 Resnet50	3rd Gen	73%
1000	Base2 Resnet50	Combined	70%

Table 3. Results without Adversarial Attack 1000 samples

Samples	Model	Test Set	Accuracy
2000	Base1 VGG16	2nd Gen	94%
2000	Base1 VGG16	3rd Gen	92%
2000	Base1 VGG16	Combined	90%
2000	Base1 Resnet50	2nd Gen	91%
2000	Base1 Resnet50	3rd Gen	89%
2000	Base1 Resnet50	Combined	90%
2000	Base2 VGG16	2nd Gen	91%
2000	Base2 VGG16	3rd Gen	88%
2000	Base2 VGG16	Combined	89%
2000	Base2 Resnet50	2nd Gen	89%
2000	Base2 Resnet50	3rd Gen	89%
2000	Base2 Resnet50	Combined	87%

Table 4. Results without Adversarial Attack 2000 samples

Samples	Model	Test Set	Accuracy
2000	Base1 VGG16	2nd Gen	98%
3000	Base1 VGG16	3rd Gen	95%
3000	Base1 VGG16	Combined	97%
3000	Base1 Resnet50	2nd Gen	94%
3000	Base1 Resnet50	3rd Gen	93%
3000	Base1 Resnet50	Combined	94%
3000	Base2 VGG16	2nd Gen	95%
3000	Base2 VGG16	3rd Gen	95%
3000	Base2 VGG16	Combined	92%
3000	Base2 Resnet50	2nd Gen	93%
3000	Base2 Resnet50	3rd Gen	92%
3000	Base2 Resnet50	Combined	93%

Table 5. Results without Adversarial Attack 3000 samples

Model Model	FGSM Blackbx	FGSM Whitebx	CWL2 Blackbx	CWL2 Whitebx
Base1 VGG16	8.7%	3.5%	22.2%	0.9%
Base1 Resnet50	15.3%	8.2%	4.3%	0.7%
Base2 VGG16	5.4%	0.0%	15.9%	0.0%
Base2 Resnet50	12.9%	0.3%	7.1%	0.1%

Table 6. Adversarial Attack 3000 samples

Base1 models are the implementation of the deepfake video detectors using Temporal Characteristics discussed in Section 2.3.1.Base2 models are implementations of the deepfake video detectors using the facial artifacts of the eye discussed in Section 2.3.2. Samples listed in tables are training samples equally drawn from second-generation deepfake dataset Celeb-df [29] and Facebook DFDC [11]. The three test sets were of 100 samples drawn either entirely from Celeb-df or entirely from Facebook DFDC or equally drawn from both of them.

A general trend was observed that the model accuracy of both the families increased with an increase in sample size. The maximum jump in accuracy was seen when the

sample size was doubled from 1000 samples to 2000 samples. Models created with Resnet50 [20] trailed in accuracy compared to models created with VGG16 [40]. Also, models using facial artifacts of the eye trailed in accuracy constantly when compared to models using the entire video leveraging the temporal characteristics. Also, higher accuracy was seen when models were tested using samples only from Celeb-df when compared to both, the test set comprised of samples only from Facebook DFDC and a combination of equally drawn samples.

When tested against perturbations a VGG based models were better compared to the Resnet models against Carlini Wagner L2 Norm Attacks (CWL2) in black-box situations. Resnet models outperformed VGG based models when tested against Fast Gradient Sign Method Attacks (FGSM) attacks in black-box situations. White-box attacks were highly effective against both the model families and as expected outperformed black-box attacks. Temporal models generated better accuracy compared to models leveraging Facial Artefacts of the Eye due to taking into consideration the entire frame.

7. Inferences from Discrepancies in Results

Due to hardware restrictions put in place to ensure the condition of being able to run on commodity hardware we trained both the baseline model family implementations on a maximum of 3000 samples drawn equally from two generations of datasets. Even after 3000 samples, we were not able to reach the accuracy mentioned in the baseline papers for deepfake video detectors leveraging temporal characteristics [19] and deepfake video detectors using the facial artifacts of the eye [28].

This variation can be attributed to the training data and test data as both the baseline papers generate their own training data using deepfake video generator tools discussed in Section 2.1. The absence of bench-marking datasets is thus enshrined in the result of this study. Also, we did not use the entirety of the datasets as this would violate commodity hardware restriction as specialized hardware would be needed to process the complete datasets comprising of 100,000 samples in Facebook DFDC [11] and 6000+ samples in Celeb-Df [29].

The adversarial attack result is similar to the results reported by Gandhi *et al*. [15]. The low accuracy of both implemented families depicts the low robustness of deepfake video detectors against adversarial attacks, both in black-box and white-box settings.

8. Associated Legal and Ethical Issues

Deepfake video imagery can be excessively deleterious to both humans and organizations. Surprisingly we did not find specialized laws created against the creation, usage, or deployment of such imagery in the United States of America. Though Federal and State laws exist to combat the misuse of Artificial Intelligence (AI), such laws are catered to misuse of AI, peculiarly in the matters of Defense and Homeland Security [35].

We try to study deepfake video detectors leveraging white-hat methodologies. This research and all referenced literature and methods adhere strictly to the ACM code of ethics [3].

9. Conclusion

In this study, we tried to demonstrate the necessity of a bench-marking methodology to compare two main families of deepfake video detectors. Discrepancies were found due to the nature of hardware and datasets used to train and test the models. We also tested the robustness of these models using both black-box and white-box adversarial attacks. The adversarial attacks test found that both the model families were extremely susceptible to adversarial attacks. We predict the future direction of research to move forward in the direction of making deepfake video detectors more robust against perturbations.

References

[1] Henry Ajder. Deepfake threat intelligence: a statistics snapshot from june 2020. Sensity.ai, June 2020. https://sensity.ai/deepfake-threat-intelligence-a-statistics-snapshot-f Accessed October 06, 2020.

[2] AMD. Amd ryzen processors for business. Online, April 2020. https://www.amd.com/en/processors/ryzen-processors-laptop-business?gclid=Cj0KCQiA2af-BRDzARIsAIVQUOfSFO_u1TaW3Cxt1LXevsi7iv5Lkyi4nK4ra0MlP0X0efqoj-p4WBcaAqdUwcB, Accessed December 01, 2020.

[3] Association for Computing Machinery. ACM Code of Ethics and Professional Conduct. ACM, New York, 2018. Webpage: https://www.acm.org/code-of-ethics.

[4] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K. Nayar. Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph.*, 27(3):1–8, Aug. 2008.

[5] Google AI Blog, September 2019. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html, Accessed November 9, 2020.

[6] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. AISec '17, page 3–14, New York, NY, USA, 2017. Association for Computing Machinery.

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017.

[8] C. Chan, S. Ginosar, T. Zhou, and A. Efros. Everybody dance now. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5932–5941, 2019.

[9] Nvidia Corporation. Geforce gtx 16 series graphics card. Online, June 2019. https://www.nvidia.com/en-us/geforce/graphics-cards/gtx-1660-ti/, Accessed December 01, 2020.

[10] Kevin Dale, Kalyan Sunkavalli, Micah K. Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. Video face replacement. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, SA '11, New York, NY, USA, 2011. Association for Computing Machinery.

[11] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020.

[12] faceswap GAN github, November 2019. https://github.com/shaoanlu/faceswap-GAN, Accessed November 9, 2020.

[13] faceswap github, November 2019. https://github.com/deepfakes/faceswap, Accessed November 9, 2020.

[14] FakeApp, November 2019. https://www.malavida.com/en/soft/fakeapp/, Accessed November 9, 2020.

[15] Apurva Gandhi and Shomik Jain. Adversarial perturbations fool deepfake detectors, 2020.

[16] DeepFaceLab github, November 2019. https://github.com/iperov/DeepFaceLab, Accessed November 9, 2020.

[17] DFaker github, November 2019. https://github.com/dfaker/df, Accessed November 9, 2020.

[18] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[19] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

[22] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019.

[23] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019.

[24] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[25] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Trans. Graph.*, 37(4), July 2018.

[26] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *CoRR*, abs/1812.08685, 2018.

[27] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3697–3705, 2017.

[28] Y. Li, M. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, Dec 2018.

[29] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics, 2020.

[30] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 700–708, Red Hook, NY, USA, 2017. Curran Associates Inc.

[31] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 405–415, Red Hook, NY, USA, 2017. Curran Associates Inc.

[32] Paarth Neekhara, Shehzeen Hussain, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples, 2020.

[33] Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. Deep learning for deepfakes creation and detection: A survey, 2020.

[34] Y. Nirkin, Y. Keller, and T. Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7183–7192, 2019.

[35] Law Library of Congress. Regulation of artificial intelligence: The americas and the caribbean, Jul 24 2020. https://www.loc.gov/law/help/artificial-intelligence/americas.php, Accessed November 22, 2020.

[36] Dang Pham and Tuan M. V. Le. Auto-encoding variational bayes for inferring topics and visualization, 2020.

[37] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[38] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *CoRR*, abs/1901.08971, 2019.

[39] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces, 2018.

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[41] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. What makes tom hanks look like tom hanks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3952–3960, 2015.

[42] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4), July 2017.

[43] Shahroz Tariq, Sangyup Lee, and Simon S. Woo. A convolutional lstm based residual network for deepfake video detection, 2020.

[44] Taylor Telford. The world is not yet ready for deepnude': Creator kills app that uses ai to fake naked images of women. Washington Post, June 2019. https://www.washingtonpost.com/business/2019/06/28/the-world-is-not-yet-ready-deepnude-creator-kills-app-that-uses-ai-fake-naked-images-women/, Accessed October 06, 2020.

[45] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 38(4), July 2019.

[46] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016.

[47] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. *International journal of computer vision*, 128(7):1867–1888, 2020.

[48] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.

[49] Walt Woods, Jack Chen, and Christof Teuscher. Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nature Machine Intelligence*, 1(11):508–516, Nov 2019.

[50] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, 2019.

[51] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9458–9467, 2019.