Evaluating Robustness of Sequence-based Deepfake Detector Models by Adversarial Perturbation

Shaikh Akib Shahriyar
Rochester Institute of Technology
Rochester, New York, USA
as8751@rit.edu

Matthew Wright
Rochester Institute of Technology
Rochester, New York, USA
matthew.wright@rit.edu

ABSTRACT

Deepfake videos are getting better in quality and can be used for dangerous disinformation campaigns. The pressing need to detect these videos has motivated researchers to develop different types of detection models. Among them, the models that utilize temporal information (i.e., sequence-based models) are more effective at detection than the ones that only detect intra-frame discrepancies. Recent work has shown that the latter detection models can be fooled with adversarial examples, leveraging the rich literature on crafting adversarial (still) images. It is less clear, however, how well these attacks will work on sequence-based models that operate on information taken over multiple frames. In this paper, we explore the effectiveness of the Fast Gradient Sign Method (FGSM) and the Carlini-Wagner L_2-norm attack to fool sequence-based deepfake detector models in both the white-box and black-box settings. The experimental results show that the attacks are effective with a maximum success rate of 99.72% and 67.14% in the white-box and black-box attack scenarios, respectively. This highlights the importance of developing more robust sequence-based deepfake detectors and opens up directions for future research.

CCS CONCEPTS

• **Applied computing** → **Computer forensics**; • **Computing methodologies** → **Computer vision**; • **Human-centered computing** → Collaborative and social computing theory, concepts and paradigms.

KEYWORDS

Deepfake detection, adversarial attacks, sequence-based model, adversarial perturbation

ACM Reference Format:
Shaikh Akib Shahriyar and Matthew Wright. 2022. Evaluating Robustness of Sequence-based Deepfake Detector Models by Adversarial Perturbation. In *Proceedings of the 1st Workshop on Security Implications of Deepfakes and Cheapfakes (WDC '22), May 30, 2022, Nagasaki, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3494109.3527194

WDC '22, May 30, 2022, Nagasaki, Japan

ACM ISBN 978-1-4503-9178-8/22/05…$15.00
https://doi.org/10.1145/3494109.3527194

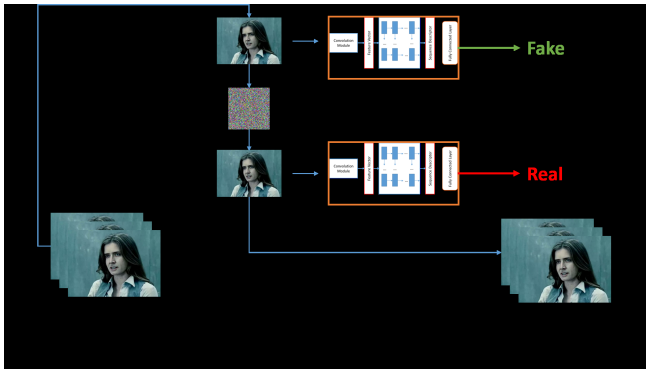

Figure 1: The process of crafting adversarially perturbed deepfakes, in which a perturbation is added to each frame.

1 INTRODUCTION

Deepfakes are a form of synthetic media in which a target individual's likeness is swapped with someone else or manipulated to move and speak as the creator desires. Although there are positive uses of this technology, it has been used to make non-consensual pornography [9, 36] and are a serious threat for spreading misinformation online [18]. To combat the spreading of misinformation via deepfakes, Facebook, Twitter, and Microsoft joined forces to remove deepfakes from online platforms in 2020 [4, 37].

Furthermore, researchers have come up with various deep-learning-based solutions to detect deepfakes. CNN-based deepfake detection methods usually try to find discrepancies in each frame of the deepfake video independently from the other frames [1, 8, 11]. Although these approaches are sound, they do not consider the temporal coherence of the frames. Researchers have since introduced sequence-based models such as Conv-LSTM, which can detect the inter-frame temporal inconsistencies in a deepfake video and thus perform significantly better than CNN-based models [7, 15, 31].

Naturally, since these detection models are based on machine learning, they are vulnerable to being fooled by adversarial examples. Researchers have studied the vulnerability of CNN-based deepfake detectors and found them to fare poorly [5, 13, 16]. A general process to craft adversarial deepfakes is shown in Fig. 1. The sequence-based models are relatively more complex, so it is unclear how well the same techniques will fare. Since the sequence-based models have greater detection accuracy than the CNN-based ones [7, 15, 31], they are more likely to be deployed. Thus, assessing the vulnerabilities of these sequence-based models against adversarial perturbations is crucial in making them robust against future adversarial attacks.

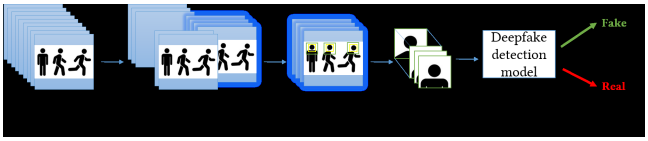

Figure 2: An overview of deepfake video detection. Frames are sampled from the raw input video and passed to a face-detection layer that crops the faces from the frames with extra margins and feeds them to the deepfake detection model for inference.

In this paper, we evaluate the effectiveness of the Fast Gradient Sign Method (FGSM) [14] and Carlini-Wagner L_2-norm (CW-L_2) attacks [6] in both white-box and black-box settings to fool sequence-based deepfake video detector models. The experimental results show that the attacks are effective against these models, with a maximum success rate of 99.72% and 67.14% for white-box and black-box attack scenarios, respectively. We note that our black-box attacks rely entirely on transferability, and they do not require any queries of the model, which could hinder real-world attacks. These findings highlight the importance of building defenses against adversarial perturbation for sequence-based detectors, and also opens up future directions for this research area.

2 BACKGROUND

2.1 Input Video Processing

The approach of processing each frame from a video and classifying them as real and fake is shown in Fig. 2. Deepfake detection models generally take a set of frames as input rather than the complete video. The detection model can have a separate face detection layer that detects faces in the frames and crops them with extra margins. Otherwise, the face detection and cropping can be achieved via independent face detection libraries, such as dlib [20], which preprocess the input video and feeds the cropped region to the model.

2.2 CNN-based Detectors

There are several prominent detection models that use CNN-based architectures for frame-level classification [42]. Among them, XceptionNet [8] and MesoNet [1] are among the most effective ones, with 95% and 84% reported accuracy, respectively.

2.3 Sequence-based Detectors

Videos have a time continuity that can be disrupted when the frames are manipulated by any perturbation. To potentially detect disruptions caused by deepfake generation, sequence-based models add some sequence processing layers, e.g. LSTM units, that take inputs from convolutional layers are used for image-level feature extraction. The Conv-LSTM model proposed by Güera and Delp [15] was the first sequence-based model for deepfake detection. FacenetLSTM [31] uses a similar approach, but with a pre-trained FaceNet model [30] for the convolutional layers.

Further research has also taken place in advancing the capabilities of sequence-based deepfake detectors [42]. A comparison of the CNN-based and sequence-based models in terms of detection accuracy on the FaceForensics++ dataset is shown in Tab. 1. The sequence-based models generally outperform the CNN-based ones and are thus more likely to be deployed. Thus, evaluating their effectiveness against adversarial perturbation and finding their vulnerabilities is becoming increasingly important.

2.4 Adversarial Examples

Fig. 1 depicts the process of generating white-box adversarial examples for deepfake videos. First, frames are extracted from the input videos as explained in §2.1. Each extracted frame is then perturbed by using an attack algorithm based on the detector model. Using the FGSM attack for example, an input frame is fed to the detector. The detector correctly classifies it as fake, but the attack utilizes the loss of the detector network and backpropagates the gradient of the network back to the input frame. The attack then uses this to determine the direction of perturbation to apply to the input frame to make the model classify it as a real frame. This process is performed on every input frame, and the resulting frames are classified as real.

In case of a black-box approach, the assumption is that the attacker would not have prior access to the model architecture. A typical approach in this case is to have the attacker to perform queries to the target model and get the classification probabilities to craft effective adversarial examples without any gradient information. This approach, however, can require a huge number of queries and can be difficult in realistic scenarios, where the attacker might have a limit on the number of queries they can perform. Another solution is to perform black-box attacks based on the *transferability* of adversarial examples. Researchers have found that the adversarial examples generated for a reference model are likely to remain effective for the target model, even transferring between different families of models and models trained on different data [24, 27].

3 RELATED WORK

In this section, we take a look at the existing research in crafting adversarial examples (e.g. adversarial deepfake videos), and also dive into the vulnerability analysis of different deepfake detection methods by other prominent works.

3.1 Adversarial Video Generation

Significant work has been done to craft adversarial videos by following different techniques [19, 23, 39, 40, 43]. The study of these techniques are valuable in exploring potential attack approaches and crafting more effective adversarial deepfake videos. Wei et al. were the first researchers to propose an adversarial video crafting

Table 1: Deepfake detection accuracy on FaceForensics++ [28], as reported in the respective papers

CNN-based Deepfake Detectors	Accuracy (%)	Sequence-based Deepfake Detectors	Accuracy (%)
XceptionNet [8]	96.71	XcepTemporal [7]	100
MesoNet [1]	84.00	Sabir et al. [29]	96.90
Amerini et al. [2]	81.61	FaceNetLSTM [31]	93.71

process [39]. The authors aim to limit the amount of perturbation to the video by using an L_1 norm across frames, meaning that the attack becomes optimized to only perturb a few select frames, and an L_2 norm within each perturbed frame. They find that the perturbations lead to misclassification of later frames due to the temporal nature of the video classifier.

Li et al. developed a targeted 3D adversarial perturbation using Generative Adversarial Network (GAN)-like architecture that works on real-time video classifiers [23]. These video classifiers use a sliding window approach to extract frames from real-time video stream; the same approach is used by the GAN to add perturbations. Although Li et al.'s work is similar to the work of Wei et al. [39], it is quite tricky to train a GAN that produces video agnostic perturbations with higher success rate at each inference.

Zajac et al. crafted adversarial videos by adding an adversarial border to each original frames of the source video [43]. While this technique has its merit, it would not be effective in crafting adversarial deepfake videos. The adversarial deepfake videos have to be visually imperceptible from the original deepfake videos to fool users into thinking that the video has not been manipulated. The borders in this technique are highly visible and clearly atypical.

All the previous work discussed above has considered crafting adversarial videos in a white-box setting. Jiang et al. proposed the first black-box video attack technique, named "V-BAD" where the prior assumption is that the adversary can only query the target model for class labels or probabilities [19]. The average number of queries needed ranges between 3400 to 8400. Wei et al. developed a black-box heuristic-based algorithm to find out the importance of each frame in an input video and also to locate the salient region of the input frames and perform a targeted attack [40]. Although the reported method achieved 100% attack success rate, it uses huge number of queries such as, 190,000 for a targeted attack and 14,000 for an untargeted attack.

Although, these black-box techniques are good, they both involve large number of queries to the victim model. For our proposed work, we will not allow the attacker query access to the model, making these techniques impossible to use.

3.2 Vulnerability of Deepfake Detectors

Three research groups have investigated the vulnerability of CNN-based deepfake detector models to adversarial examples. Gandhi et al. tested the robustness of two vanilla CNN-based image classifiers in a deepfake image detection task [13]. They chose the VGG and ResNet architectures and trained them on a custom dataset to detect deepfake images. They then perturbed the images using FGSM [14] and CW-L_2 [6] attacks. They reported that the deepfake image detector's accuracy dropped from 95% to under 27%. Carlini et al. [5] proposed multiple white-box and black-Box attacks on two CNN-based deepfake image classifiers, from Wang et al. [38] and Frank et al. [12]. The white-box attacks on those classifiers reduced their accuracy to almost 0% and the black-box attacks resulted in the reduction of the area under the ROC curve (AUC) from 0.95 to 0.22. Both of these works considered only deepfake image classifiers for their study, while not studying deepfake video detectors nor temporal models.

Hussain et al. [16] tested the robustness of XceptionNet [8] and MesoNet [1], two of the best performing CNN-based deepfake video detectors. They used the IGSM attack [22] for a white-box attack and NES [17, 41] for robust white-box [3] and black-box attacks. They report that the average success rate of their white-box attacks was 99.85% for XceptionNet and 98.15% for MesoNet. In the case of black-box attacks on raw format videos, the average success rates were 97.04% for XceptionNet and 86.70% for MesoNet. The authors also considered a pre-trained 3D-CNN model for evaluating a form of sequence-based detector. Their attacks were less successful than those on XceptionNet and MesoNet, in both white-box and black-box settings. Although this work evaluates deepfake video detectors, the 3D-CNN attacked here only learns temporally local features, while CNN+RNN/LSTM based models learn temporally global features. Better performing deepfake detectors utilize the latter architecture, making it important to establish that CNN+RNN/LSTM-based models are vulnerable to adversarial perturbation. To the best of our knowledge, no other work has been done to test the robustness of these CNN+RNN/LSTM-based models which leaves open a potential research area to explore.

4 SYSTEM DESIGN

The purpose of this research is to determine the robustness of sequence-based deepfake detectors to adversarial perturbations. To this end, we have designed a realistic threat model, chosen two sequence-based models as our victim models, and selected two prominent adversarial attacks that were utilized in prior literature to attack CNN-based deepfake detectors.

4.1 Victim Models

We have considered two pioneering sequence-based deep learning models that employ convolutional LSTM architectures for detecting deepfake videos. The first victim model is Conv-LSTM in which a CNN is used for feature extraction and an LSTM is used for sequence processing [15].The model works as an end-to-end deepfake video detection system. For the convolution model, they have adopted Inception V3 [32], where the fully-connected final layer is removed. The model takes each frame from the deepfake video and provides a corresponding 2048-dimensional feature vector as output. This 2048-dimensional feature vector is then fed into a 2048-wide LSTM unit with a dropout rate of 0.5. The output from this LSTM unit is then forwarded to a 512-unit fully-connected layer that also has a dropout rate of 0.5. Finally, a softmax layer is used to compute the probability of each frame being either real or fake. We implemented Conv-LSTM according to the architecture and parameters discussed by Güera and Delp [15].

FacenetLSTM, proposed by Sohrawardi et al. [31], builds upon the work of Güera and Delp. The authors replace the convolutional module (Inception V3) of the prior Conv-LSTM model with the FaceNet[30] architecture. FaceNet differs significantly from the Inception V3 architecture, as it tries to create a compact latent representation of the input face and also transforms the input face into a frontal face. We implemented and trained the model according to the architecture and parameters described by Sohrawardi et al. [31].

4.2 Threat Model

To evaluate the robustness of the victim deepfake video detector models, we have used FGSM and CW-L_2 attacks in both white-box and black-box settings. We chose these attacks based on their

Table 2: Parameters for FGSM and CW-L_2 attack

Attack	Hyperparam.	Fine-tuned Value	Search Space
FGSM	ϵ	0.03	[0.01 to 0.5]
CW-L_2	c	[100 to 10000] Search step = 5	[10^{-10} to 10^{10}]
	Learning rate	0.001	-
	Max Iteration	1000	-
	κ	200	[0 to 500]

popularity and effectiveness in attacking CNN-based models as reported in [5, 13].

The goal of the adversary in both settings is to perturb each frame of the target deepfake video to such a degree that the full video is classified as real by the victim models. Also, the adversary seeks to craft adversarial deepfake videos that are visually imperceptible from the original deepfake videos or at least minimally modified.

White-box Attack. To examine a worst-case setting and establish a baseline, we assume in this setting that the adversary has complete access to the victim model's architecture and parameters.

Black-box Attack. In case of more realistic black-box attacks, we assume the adversary has no access to the model architecture. Also, we assume the attacker cannot perform any queries to the victim model, and instead must rely on the *transferability* of samples generated against one model working on the victim model [14, 27, 33]. We created adversarial examples based on the FacenetLSTM architecture and used them to perform black-box attacks on the Conv-LSTM model, and vice versa.

5 EXPERIMENTAL DESIGN

We preprocessed the dataset, trained the victim model, and performed adversarial attacks using our local server consisting of two NVIDIA GTX 2080 GPUs in a Linux environment.

5.1 Dataset

We use the FaceForensics++ dataset developed by Rossler et al. [28]. Hussain et al. [16] evaluated the robustness of XceptionNet [8] and MesoNet [1] on this same dataset. Moreover, the FacenetLSTM was primarily trained on it [31]. The FaceForensics++ dataset consists of 1000 source videos which are manipulated by four different deepfake generation methods: DeepFakes (DF) [10], Face2Face (F2F) [35], FaceSwap (FS) [21] and NeuralTextures (NT) [34]. The mapping of training, validation and test sets are kept the same as mentioned in the Github repository of Rossler et al. [28] which include 720, 140 and 140 samples, respectively. Both of the victim models are trained on the raw training videos of this dataset.

5.2 Attack Parameters

Tab. 2 shows the attack parameters used in this study. The values of ϵ for FGSM and c and κ were chosen subjectively based on how they distorted the examples and objectively based on how they decreased the accuracy of the victim models. We have left other parameters to their respective default values as recommended in the CW-L_2 implementation from the CleverHans [26] library.

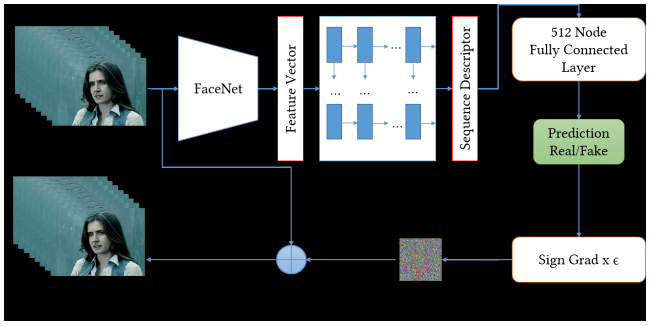

Figure 3: Adversarially perturbed frame generation using the FacenetLSTM model and the FGSM attack. Each frame is first extracted from the deepfake video and fed into the FaceNet model to generate the feature vector. The feature vector is passed all the way through the 256-node LSTM module and fully connected layer to collect the sign gradient of the loss function with respect to the input frame. The sign gradient is multiplied by the ϵ value to generate the adversarial perturbation, which in turn is added to the input frame to craft the adversarially perturbed deepfake video frames.

5.3 Crafting Adversarial Videos

Even though sequential models are more complex than CNN models, they are surprisingly easy to apply existing adversarial attacks to. An overview of the perturbation process using FGSM against the FacenetLSTM model is shown as an example in Fig. 3. The model produces a prediction for every frame of the video, and we can use this to compute the gradient of the loss. For the CW-L_2 attack, we can instead compute the gradient of the logits. The resulting perturbation is then combined with the original frame. We note that this approach is somewhat naive, in that it does not optimize jointly over multiple frames. We leave exploration of this to future work.

5.4 Evaluation Metrics

To evaluate the robustness of the victim models, we use three metrics:

Success Rate (SR).. SR is the ratio of frames that are perturbed and classified as "real" and the total number of frames in a corresponding video. A lower SR value represents a higher degree of robustness in the victim models.

Accuracy. Classification accuracy of each victim model is evaluated on the unperturbed test set, FGSM-perturbed test set, and CW-L_2-perturbed test set.

Mean L_∞ Distortion. We calculate the average L_∞ distortion between the original frames and adversarial frames. The pixel values of each frame are scaled in the range [0,1], so the maximum value of L_∞ is 1 if the adversarial frame is completely distorted.

6 EXPERIMENTAL RESULTS

Following §4, we have implemented the victim models and crafted adversarial examples from the specified dataset. The Conv-LSTM model achieved an accuracy of 81.3%, and the FacenetLSTM model achieved an accuracy of 84.5% after training on the unperturbed FaceForensics++ training set. Both of models were trained for 10 epochs and converged successfully. The average training time for

Table 3: Model accuracy and attack success rate

Sn	Model	Unperturbed Accuracy	White-box Attack				Black-box Attack			
			FGSM		CW-L_2		FGSM		CW-L_2	
			Accuracy	SR	Accuracy	SR	Accuracy	SR	Accuracy	SR
1	Conv-LSTM	81.3%	14.8%	72.31%	8.3%	99.72%	44.7%	21.73%	38.4%	63.82%
2	FacenetLSTM	84.5%	20.9%	68.23%	13.5%	98.83%	53.5%	33.89%	28.7%	67.14%

each epoch was 8.5 and 11.3 hours for Conv-LSTM and FacenetLSTM, respectively.

Tab. 3 shows the performance of these models in both white-box and black-box settings. Recall that we performed the black-box attack in a cross-model manner, such that the adversarial examples used for attacking the Conv-LSTM model were generated using the FacenetLSTM architecture and vice-versa.

Accuracy and Success Rate. The white-box attacks reduced the accuracy of the victim models to 8.3%-20.9%, while the black-box attacks reduced their accuracy to 28.7%-53.5%. The attack success rate (SR) shows similar patterns. As expected, the CW-L_2 attack outperforms FGSM. The results are in line with those achieved by Gandhi and Jain on CNN-based deepfake image detectors [13].

Distortion. Tab. 4 shows the mean distortion results on the Faceforesnsics++ dataset. As our attack model perturbed each test video using both FGSM and CW-L_2 for both of the victim models, we have calculated the mean L_∞ distortion for each type of deepfake dataset. The average L_∞ distortion ranged from 0.0548 to 0.0613. These distortion values are higher than those reported by Hussain et al. [16], which ranged from 0.04-0.047. We note that our attack needs to fool both convolutional and LSTM layers, which explains the greater distortion required. Gandhi et al. and Carlini et al. did not use the FaceForesnics++ dataset for training and testing their victim models [5, 13], and thus we have excluded their work while comparing distortion values.

Perturbation vs. Imperceptibility. We have explored the search space of the FGSM and CW-L_2 attack hyperparameters (Tab. 2) to understand the trade off between perturbation and imperceptibility. For FGSM, we perturbed the frames with different ϵ values ranging from 0.01 to 0.5. As the perturbation nears 0.5, model accuracy drops close to 1%. Fig. 4 shows the relationship between the amount of perturbation and subjective imperceptibility in case of FGSM attack. The frames are collected randomly from four different types of deepfake videos available in the FaceForensics++ test set. We perturbed the input frames according to different ϵ values in the range of 0.01 to 0.5, with a step size of 0.01. The victim models classify all frames with $\epsilon \geq 0.1$ as *Real*. We can see that the input frames become more distorted as the amount of perturbation increases. The L_∞ value under each frame gives us an idea about the amount of distortion present at each perturbed frame. It grows significantly as ϵ grows from 0.03 to 0.1, as the perturbation becomes visible, and against as ϵ grows from 0.1 to 0.3, as the perturbation becomes highly noticeable.

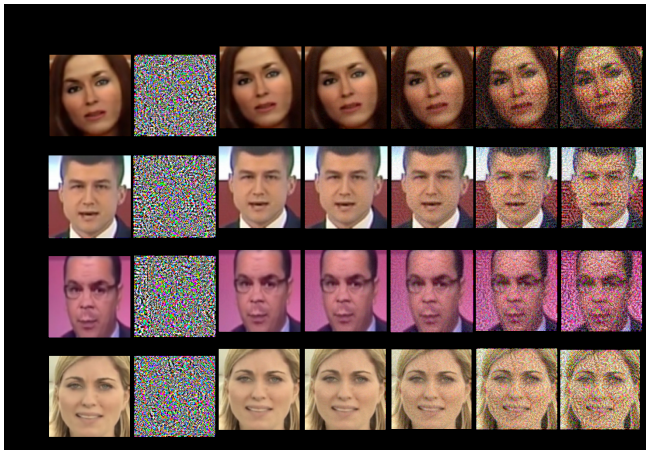

Figure 4: Adversarially perturbed frames extracted from randomly selected deepfakes from the FaceForensics++ test set. The input frames are classified as *Fake* by the victim models. The perturbed frames generated according to different ϵ values using FGSM are shown in each row. Perturbed frames whose $\epsilon > 0.01$ are labeled as *Real* by the victim models. The distortion increases proportionally with ϵ.

Table 4: L_∞ distortion results. Datasets from FaceForensics++ [28]: DeepFakes (DF) [10], Face2Face (F2F) [35], FaceSwap (FS) [21] and NeuralTextures (NT) [34]

Sn	Dataset	Conv-LSTM Distortion L_∞		FacenetLSTM Distortion L_∞	
		FGSM	CW-L_2	FGSM	CW-L_2
1	DF	0.074	0.069	0.083	0.072
2	F2F	0.067	0.052	0.071	0.043
3	FS	0.057	0.055	0.049	0.065
4	NT	0.045	0.058	0.042	0.039
Average		0.0607	0.0585	0.0613	0.0548

For the CW-L_2 attack, we have performed similar exploration and found that the upper bound of the search space for c can be extended to 100000 and κ can be increased to 500 to generate more successful adversarial frames. The performance of the victim models also drops significantly (close to 1% accuracy), with a similar trade-off between the amount of distortion and imperceptibility of the adversarial frames in case of CW-L_2 attack.

7 DISCUSSION AND LIMITATIONS

Our experimental results validate our attack model and exposes the vulnerability of sequence-based deepfake detectors. There are

some limitations to our work, however. First, we note that our initial training of the FacenetLSTM model on the Faceforensics++ dataset achieved an accuracy of 84.5%, which is significantly lower than the accuracy of 93.71% reported by Sohrawardi et al. [31]. Further fine-tuning may be required. Second, there are other attack algorithms that may be more effective, such as PGD [25] or the more sophisticated video attack of Wei et al. [39]. We selected two well-established attack algorithms for this initial exploration, but identifying more effective approaches will be important for future work. Also, we have not tested the effect of different pre-processing steps on the perturbed video frames such as lossy or lossless compression. Finally, we have not tested on defenses that could be applied, such as adversarial training among many options. Exploring the space of defense designs as applied to deepfake detection is a critical avenue for future work.

8 CONCLUSION

It is quite challenging to detect deepfakes, as more realistic deepfake generation methods are being developed all the time. Sequence-based models offer state-of-the-art detection performance, but our experiments indicate that they are highly vulnerable to adversarial perturbations, including transferability-based black-box attacks. Developing defenses against these attacks will be critical for trustworthy deepfake detection systems.

ACKNOWLEDGMENTS

This material is based upon work supported in part by the John S. and James L. Knight Foundation and the National Science Foundation under Award No. 2040209.

REFERENCES

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. Mesonet: A Compact Facial Video Forgery Detection Network. In *IEEE WIFS*.
[2] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. 2019. Deepfake Video Detection Through Optical Flow based CNN. In *CVPR Workshops*.
[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. Synthesizing Robust Adversarial Examples. In *ICML*.
[4] Tom Burt. 2020. *New Steps to Combat Disinformation.* https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/. Accessed = March 1, 2021.
[5] Nicholas Carlini and Hany Farid. 2020. Evading Deepfake-Image Detectors with White-and Black-Box Attacks. In *CVPR Workshops*.
[6] Nicholas Carlini and David Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy (SP)*. IEEE.
[7] Akash Chintha, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. 2020. Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection. *IEEE JSTSP* 14, 5 (2020).
[8] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *CVPR*.
[9] Samantha Cole. 2017. *AI-Assisted Fake Porn Is Here and We're All Fucked.* https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn. Accessed = March 25, 2021.
[10] DeepFaceLab. 2020. *DeepFaceLab.* https://github.com/iperov/DeepFaceLab. Accessed = March 2, 2021.
[11] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. 2019. The Deepfake Detection Challenge (DFDC) Preview Dataset. *arXiv preprint arXiv:1910.08854* (2019).
[12] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. *ICML* (2020).
[13] A. Gandhi and S. Jain. 2020. Adversarial Perturbations Fool Deepfake Detectors. In *IJCNN*.
[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. *ICLR* (2015).
[15] D. Güera and E. J. Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In *AVSS*.
[16] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. 2021. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. In *IEEE/CVF WACV*.
[17] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box Adversarial Attacks with Limited Queries and Information. *ICML* (2018).
[18] Charlotte Jee. 2020. *An Indian Politician is using Deepfake Technology to Win New Voters.* https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/. Accessed = March 2, 2021.
[19] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. 2019. Black-box Adversarial Attacks on Video Recognition Models. In *ACM MM*.
[20] Davis E. King. 2017. *dlib C++ Library.* http://dlib.net/. Accessed = March 25, 2021.
[21] Marek Kowalski. 2020. *Faceswap.* https://github.com/MarekKowalski/FaceSwap/. Accessed = March 2, 2021.
[22] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial Examples in the Physical World. *CVPR* (2016).
[23] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy-Chowdhury, and Ananthram Swami. 2019. Stealthy Adversarial Perturbations Against Real-Time Video Classification Systems. (2019).
[24] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2017. Delving into Transferable Adversarial Examples and Black-box Attacks. In *ICLR*.
[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.
[26] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. 2018. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. *arXiv preprint arXiv:1610.00768* (2018).
[27] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. *arXiv preprint arXiv:1605.07277* (2016).
[28] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to Detect Manipulated Facial Images. In *ICCV*.
[29] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. *Interfaces (GUI)* 3, 1 (2019).
[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A Unified Embedding for Face Recognition and Clustering. In *CVPR*.
[31] Saniat Javid Sohrawardi, Akash Chintha, Bao Thai, Sovantharith Seng, Andrea Hickerson, Raymond Ptucha, and Matthew Wright. 2019. Poster: Towards Robust Open-World Detection of Deepfakes. In *ACM CCS*.
[32] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *CVPR*.
[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing Properties of Neural Networks. *ICLR* (2014).
[34] Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis using Neural Textures. *ACM Transactions on Graphics (TOG)* 38, 4 (2019).
[35] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time Face Capture and Reenactment of RGB Videos. In *CVPR*.
[36] James Vincent. 2019. *New AI Deepfake App Creates Nude Images of Women.* https://www.theverge.com/2019/6/27/18760896/deepfake-nude-ai-app-women-deepnude-non-consensual-pornography. Accessed = March 3, 2021.
[37] Kurt Wagner. 2020. *Twitter will label, remove deepfake videos under new policy.* https://fortune.com/2020/02/04/twitter-deepfake-videos/. Accessed = March 1, 2021.
[38] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-generated Images are Surprisingly Easy to Spot... for Now. In *CVPR*, Vol. 7.
[39] Xingxing Wei, Jun Zhu, Sha Yuan, and Hang Su. 2019. Sparse Adversarial Perturbations for Videos. In *AAAI*.
[40] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang. 2020. Heuristic Black-box Adversarial Attacks on Video Recognition Models. 34, 07 (2020).
[41] Daan Wierstra, Tom Schaul, Jan Peters, and Juergen Schmidhuber. 2008. Natural Evolution Strategies. In *IEEE Congress on Evolutionary Computation*.
[42] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. 2021. A Survey on Deepfake Video Detection. *IET Biometrics* 10, 6 (2021).
[43] Michał Zajac, Konrad Zołna, Negar Rostamzadeh, and Pedro O Pinheiro. 2019. Adversarial Framing for Image and Video Classification. In *AAAI*.