Reliable Facial Forgery Detection

Dr. Simmi Dutta[1], Aditya Sharma[2], Ruban Nazir Shah[3], Harsharan Singh Raina[4]

[1,2,3,4] *Department of Computer Engineering, Government College of Engineering and Technology, Jammu*

Abstract— **The free access to large-scale public databases, alongside the fast progress of deep learning techniques, especially Generative Adversarial Networks, have led to the generation of very realistic fake contents which can be threatening and have various implications in today's world where enforcing fake news is pretty simple. This survey provides a radical review of techniques for manipulating face images including Deep Fake methods, and methods to detect such manipulations. These techniques are, face synthesis, face identity swap, facial attributes manipulation, and countenance manipulation. For each manipulation type, we offer details regarding manipulation techniques, existing public data-bases, and key benchmarks for technology evaluation of faux detection methods, including a summary of results from those evaluations.**

Index Terms— **Deepfakes, FaceForensics, Resnet, XceptionNet**

I. INTRODUCTION

Misinformation present online has always been a security concern, but with the sudden advent of high quality facially manipulated videos, we are at the cusp of an era where the actual presence of people in visual media comes into question.

Presence of facially manipulated images online is nothing new, but with the help of deep learning techniques like generative adversarial networks, GANs we are now able to create extreme-ly high quality Forged videos. Deepfakes[13]is a type of these facially manipulated videos which are created using deep learning techniques where the face of an individual in a video or image is replaced with facial features of someone else. Deepfakes[13] are getting more realistic by the day and are now at a point where they cannot be reliably distinguished by the naked eye, thus their very presence online can lead to dire consequences.

To combat this issue reliably various forgery detection techniques and models have been proposed and many large scale datasets have also been contributed like FaceForensics++[14],Deep Fake Detection[7], DFDC[8], Deepfake-TIMIT[13] which contain diverse sets of forged visual media.

FaceForensics++[14], in particular, have claimed very promising results in the paper with an average accuracy of 91.99% using a fine-tuned XceptionNet[16]. However, while testing the model on random deepfakes available online, particularly Youtube, we found the accuracy of the model to be much lower than claimed accuracy on paper.

With this paper we intend to make the following contributions:

1) Identifying issues present in current state-of-the-art forgery detection dataset FaceForensics++.

2) Highlight the need for a constantly updated open-source data-set to make deepfake detection reliable.

3) Improving the dataset to make a more generalised detection model.

II. RELATED WORK

XceptionNet[16]: It is a Deep Learning Algorithm that perceives the manipulation of the facial attributes in videos. To solve the issues such as identity theft and problems posed to the biometric system, the researchers from the Technical University of Munich developed a deep learning algorithm which identified the forged videos on the internet with potential face swaps. The researchers collected a dataset which contained over 1000 videos that had face swaps and their original versions. The database that was created with these videos contained over half a million images of manipulated faced. After collecting the dataset, a deep learning neural network model was trained to understand and differentiate between the manipulated video and the original video.

Resnet[17]: ResNet is a convolutional neural network comprising of 18 layers. The ResNet model is smaller than the XceptionNet model and a pre-trained version of the network, trained on more than a million images from ImageNet database can be loaded on it. The pre-trained network classifies images into 1000 objects categories which assists the network to learn rich representations of features for a wide range of images.

FaceForensics++[14] : It is a forensic dataset which contains 1000 original videos with facial manipulation. The methods which have been used for facial manipulation are- Deepfake[13], Face2Face, FaceSwap[11] and NeuralTextures. There are a total of 1000 pristine and 4000 forged videos

Some other Deepfake Datasets-

The DeepFake-TIMIT[13]: dataset includes640 DeepFake videos generated with face swap-GAN and based on the Vid-TIMIT dataset.

DFD[7] : The Google/Jigsaw DeepFake detection dataset has 3,068 DeepFake videos generated based on 363 original videos of 28 consented individuals of various genders, ages and ethnic groups.

DFDC[8]: The Facebook DeepFake detection challenge dataset is part of the DeepFake detection challenge, which has 4,113 DeepFake videos created based on 1,131 original videos of 66 consented individuals of various genders, ages and ethnic groups.

III. BENCHMARKING

We evaluated the pre-trained XceptionNet model trained on

FaceForensics++ dataset as it out-performed all the other baseline models present in the paper. We chose frame-level AUC scores or area under the ROC curve as a metric of overall performance, this allows us to have a fair analysis of the model without the need of calibrating the model for different datasets.

Our benchmarking involved validating XceptionNet[16]across 20% data of large scale deep fake detection datasets, namely Deep Fake Detection, DFDC, Deepfake-TIMIT. For analysing its real-life performance we made a custom WEB dataset of diverse keyframes sourced from various high-quality deepfakes present across the internet

	Train	Test
Real	3384	1480
Fake	3646	2768

Table 1: Number of images in WEB Dataset with their distribution

1.WEB DATASET

We created this dataset to be representative of the average high-quality deepfakes[13] encountered online. We downloaded 50 real videos and 50 high quality deepfake unrelated videos from vari-ous websites. During the selection procedure, we made sure to select keyframes containing diverse faces and different back-grounds and lighting conditions to emulate a real-life scenario.

We assume that these deepfakes[13] are created using multiple forgery techniques cascaded on top of each other and thus prove difficult in being detected by people. We also employed simple augmentation techniques to expand the dataset to combat overfitting issues, the techniques involved are:

1) random orientation change, 2) random scaling, 3) introducing distortion or noise, 4) saturation and contrast change.

In the end, we had a high-quality deepfake dataset comprising 11,278 images with a male and female ratio of 2:1 and different skin tones ranging from white, brown to black in a ratio of ap-prox. 2:1:2.

2. EXPERIMENT

In Table 2 we show different AUC scores of the XceptionNet over all Datasets previously mentioned.

DATASET	Deepfake-TIMIT	DFD	DFDC	FaceForensics++	WEB dataset
FRAME LEVEL AUC SCORES	95.1	83.4	75.3	99.7	67.3

The model fares well for old deepfake datasets like Deepfake-TIMIT[13] but struggles with relatively new large scale datasets

Table 2: XceptionNet Frame Level AUC scores on different Datasets

DFD[7] and DFDC[8]. This shows that deepfakes have been consi-derably improved in a relatively short amount of time.

The worst performance is seen when it comes to our WEB dataset with AUC scores lower than 70%. This shows that the model is not very reliable when it comes to highly diverse and quality

deepfakes present in the wild. Another thing to note is that when we tested the model on real videos scrapped from the internet, it assumed approximately a quarter of frames to be fake bringing the accuracy close to approx. 59%. Naturally, the model is not yet ready for applications.

IV. PROBLEMS WITH THE DATASET

After verifying results in Table 2 and analyzing the dataset, we made the following hypothesis:

1 FaceForensics++ Dataset focuses on introducing a large scale dataset but doesn't pay much attention to the quality of forged videos present. This leads to a high quantity of relatively low-quality deepfakes.

2 Creation of this dataset involved taking immaculate video clips and applying four different forgery techniques individually, namely FaceSwap, Deepfakes, Face2Face and NeuralTextures to create 4 fakes from a single pristine keyframe and do not simulate actual deepfakes found in the wild. Actual convincing deepfakes involve applying multiple techniques, in addition to the mentioned ones, cascaded on a single frame alone.

3 The creation of this dataset doesn't account for a large simi-larity between the training and testing images that would arise by using four images drawn from a single source. This poor distribution is the cause of super fitting issues seen in the trained model. The model could be biased to simply memorise the facial attributes of the target.

These inherent issues in the dataset limit it from training a more generalized deepfake detector. So in a more practical scenario, it is all but unreliable.

V. MAKING A MORE RELIABLE MODEL

Our hypothesis is based on the assumption that Faceforensics++ dataset is simply not up to the mark of current deepfakes found on the internet. To combat this issue we propose to introduce these high-quality deepfakes in the dataset and further enhance it.

1. Data- Preprocessing:

In Data- Preprocessing we take our WEB dataset mentioned above and combine them with 20% of FaceForensics ++ data-set. We also distribute the data much more uniformly to ensure that our model does not simply memorize the faces present in the dataset.

2. Modelling:

For modelling, we have used ResNet18[17]instead of Xception net to reduce the chances of data overfitting and to do faster iterations. This is effective since the former model is smaller than the latter.

We make use of the pre-trained ResNet18[17] model on Image-Net in all our experiments and we unfreeze the weights to be fine-tuned

on the deepfake detection task. The unfreezing of the convolu-tional layers is done to move the weights towards learning fea-tures that are more useful for algorithmic deepfake detection — artefacts, skin colour change, blur, etc. — than learning to detect

what humans would perceive as a typical set of facial features — eyes, ears, noses, etc.

The reason behind this is to enforce it to look for more features instead of just memorizing the difference between the real and fake faces. When the facial features such as eyes and ears that are extracted by ResNet18[17] are used by a fully connected classifier, that enforces the model to memorise that certain faces are associated with the label 'real,' while others are associated with the label 'fake.' To overcome this we also fine-tune the ResNet18 layers so that it takes into consideration other artificers useful in deepfake detection, such as blur or two sets of eyes appearing on a single face. With the help of these features, our fully connected classifier would be able to generalize better to faces it has never seen before.

VI. EXPERIMENTATION AND RESULTS

Training and validating the model on both datasets fetched promising results as seen by the ROC curve. The model can pick both low quality and high-quality random deepfakes with much higher accuracy and AUC score of 90+. It is also important to note that training of model on WEB dataset alone did not fetch great results where the model was only able to detect subtle image distortion in low-quality deepfakes while failing to recognise highly compressed ones. Thus introducing standard facial forgery techniques by using Faceforensics++ to the model greatly enhances it to better identify facial distortion, signalling that simply introducing random deepfakes to a model is also not enough to get a reliable solution rather we require both high quality random and low-quality standard deepfakes techniques to train our model to identify these distortions more accurately.

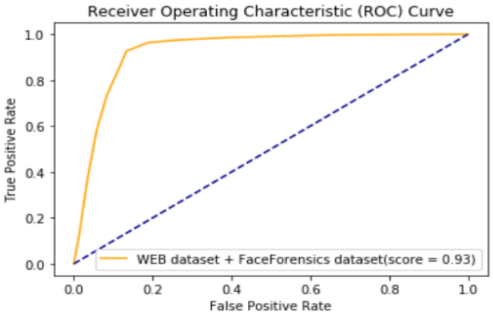

Fig 1: AUC curve of model trained on both FF+ and WEB dataset

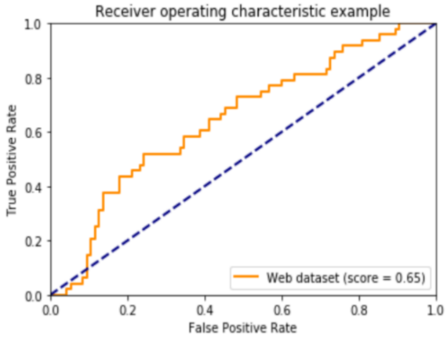

Fig 2: AUC curve of model trained on WEB dataset

VII. CONCLUSION AND FUTURE SCOPE

In this work we establish that Facial Forgery Detection issue is still far off from being solved. Although introduction of large-scale deepfake datasets are helping us to make better models, they alone are simply not enough. From our observations in thispaper, we can conclude that what we require is a large-scale open source dataset which not only contains standard deepfake creation techniques but is also constantly updated with new high quality deepfakes found on the internet. For now we must strive to continuously improve machine learning-powered deepfake detectors with newer and better data samples.

The very nature of this problem is as such that model used to detect deepfakes can be used to make even more convincing fakes thus we need to constantly update our detection techniques to match the growing manipulated images. Finding ways of identifying deepfakes ironically tends to provide those developing models used to generate them with techniques to make them more advanced. With these new techniques built into successive models, these new techniques become capable of eluding once reliable deepfake detector systems.

Although solving the issue of identity security might seem hopeless as of now, it is worth pursuing and continuous research on facial forgery will prove to be an essential foundation to the solution.

REFERENCES

[1] Zahid Akhtara, Dipankar Dasguptaa, Bonny Banerjeeb. “Face Authenticity: An Overview of Face Manipulation Generation, Detection and Recognition”, International Conference on Communication and Information Processing (ICCIP-2019)

[2] Akhtar, Zahid &Rattani, Ajita& Hadid, Abdenour&Tistarelli, Massimo). Face Recognition under Ageing Effect: A Comparative Analysis. International Conference on Image Analysis and Processing 2019.

[3] Deepfakesgithub. https://github.com/deepfakes/faceswap

[4] Faceswap.https://github.com/MarekKowalski/FaceSwap/

[5] Fakeapp. https://www.fakeapp.com/.

[6] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and SudheendraVijayanarasimhan. ”YouTube-8m: A largescale video classification benchmark”. arXiv preprint arXiv:1609.08675, 2016. 12

[7] Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Vic-tor Vorbyov, Thomas Leung, Jeremiah Childs, and ChristophBregler. “Deepfakes detection dataset by google & jigsaw”

[8] Brian Dolhansky, Russ Howes, Ben Pflaum, NicoleBaram, and Cristian Canton Ferrer. “The

deepfake detection challenge (DFDC) preview dataset”. arXiv preprintarXiv:1910.08854, 2019

[9] KyungjuneBaek, Duhyeon Bang, Hyunjung Shim. “Editable Generative Adversarial Networks: Generating and Editing Faces Simultaneously”, arXiv preprint arXiv:1807.07700

[10] Scherhag, Ulrich, Rathgeb, Christian, Busch, Christoph.“Performance variation of morphed face image detection algorithms across different datasets”, International Workshop on Biometrics and Forensics (IWBF)

[11] Ferrara, M., Franco, A., Maltoni, D.“Fast Face-swap Using Convolutional Neural Networks”, ICCV, p. 3697-3705. 2014

[12] David Güera, Edward J. Delp. “Deepfake video detection using recurrent neural networks”, IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)

[13] PavelKorshunov, Sebastien Marcel. “DeepFakes: A New Threat to Face Recognition? Assessment and Detection”, arXiv:1812.08685

[14] Andreas R ¨ossler, Davide Cozzolino, Luisa Verdoliva, Chris-tian Riess, Justus Thies, and Matthias Nießner. FaceForen-sics++: Learning to detect manipulated facial images. InICCV, 2019

[15] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAl-mageed, IacopoMasi, and Prem Natarajan. “Recurrent-convolution approach to deepfake detection-state-of-art results on faceforensics++”. arXiv preprint arXiv:1905.00582,2019

[16] François Chollet. “Xception: Deep Learning with Depthwise Separable Convolutions”. arXiv:1610.02357

[17] Kaiming He,Xiangyu Zhang,Shaoqing Ren, Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv:1512.03385v1, 10 Dec 2015