Deepfake Video Generation & Detection:Analysis

Mayur Patidar
B.Tech,
Dept. Computer Science and Engineering,
SRMIST, KTR, Chennai, India
mk5097@srmist.edu.in

Rohit R Nair
B. Tech,
Dept. Computer Science and Engineering,
SRMIST, KTR Chennai, India
rr1994@srmist.edu.in

*Dr. Kanchana M.
Associate Professor,
Dept. Computer Science and Engineering,
SRMIST,KTR, Chennai,India
kanchanm@srmist.edu.in

Abstract- In the coming years technology has advanced in so many ways that it has contributed to the cyber society with a resource that only machines alone can excel at, like art of forgery of media. This contrivance is widely known as Deepfakes. Most well documented Deep Fakes are created using Generative Adversarial Network (GANs)[10] Models that basically consists of two different Machine Learning Models playing offence and defence. These models generate and detect deepfakes until they reach a point wherein the morphing ceases detection. This algorithm/model learns to identify and generate new media with the same demography as the training set, hence developing the perfect Deep Fake media.

The changes cannot be tracked with naked eye, as modifications are done using advance features, but to design an algorithm that can automatically detect this tampering done over the media is entirely possible. This not only allows us to span our search over a single media object but also over a massive database of such mixed media. And as AI takes over on full force with automation, the more it learns, the better it gets. Over the years, new models are being introduced to generate better deep fakes hence making it harder to differentiate between legitimate and morphed media.

Keywords – Deepfake, Detection, Generation, GANs, SVM, Logistic Regression, Random Forest, Multilayer Perceptron

I. Introduction

Deepfake by definition is an Artificial Intelligence based video manipulation technique of digitally altering a person's face/body to mimic someone else with the intent of malicious use or harm via the spread of false information. In April 2018, a short one-minute video clip of former United States President Barack Obama went viral showing him discussing topics that he had never commented on before. Although the video was easily detected as a fake, by trained professionals in the IT field, it was not perceived as the same from the uneducated lower classes to the middle-class folk.

Development in AI, deep learning, machine learning and image processing technology has made deepfake videos generation very easy and realistic. In order to create deepfakes 2 auto-encoding algorithms are trained over a large collection of data while also compressing the generated data into savepoints called data points. The difference being the second algorithm does the operation over the target video. These savepoints are them swapped by both the target and source to acquire a perfect lip-to-chin effect, thereby making a successful deepfake[1]. As time goes on it is becoming more and more easier to implement and deploy this technology by the use of Apps like Fakeapp.

These models examine the facial expressions, features,contours,blood flows movements of the source and synthesize the facial features of the target generating life like facial expressions and positions with movements. Most Deepfake algorithms generally need an enormous amount of source data to train and process models in order to create a naturalistic output. Due to the abundance of images of celebrities and famous personalities it is highly common for these figures to become prime targets of this attack.

There is hence an alarming need for counteractive measures to be implemented in order to regulate and trace down fake sources of data.

II. State of art

In this work, we have analyzed different methods that are capable of distinguishing between an AI-generated fake video versus real video. We have evaluated and tested various different methods using different features , on various sets of publicly available DeepFake medias that demonstrate the efficiency and effectiveness of each algorithm in practice. In the Initial phases we would focus on validation of the model. We have compared various detection models and explored its limitations as well have found the models which transcend the limits.

Based on the observation we have also explained the basic architecture that almost every model has followed. Some of the Deepfake Detection methods and models are still in their early development stage, also multiple methods had been proposed and evaluated but only on fragments of data sets[6]. With the continuous evolution in deepfake technology, every new algorithm gives rise to a new branch of loopholes that need to be traced and neutralized iteratively.

III. Deepfake Generation

Deepfakes can be generated using various techniques and implementations. Most good Generators run on GAN’s (Generative adversarial networks) that require the participation of two different neural networks running to and against each other. This proper Police and Thief chase improves the chances of output being at realistic levels. GAN’s are an interesting Choice because they are capable of generating images as well as features automatically without the need of assistance. Simple apps like FakeApp makes it easy for even a beginner to use and exploit the frameworks of deepfakes provided if he/she knows how the framework works. The app only requires you to provide the target and source image of data, not only does it automatically download the source and target data, the software also automatically pre-processes and filters the images automatically. There are many other methods and algorithms that work in the similar way like Face Forensics++, Face2Face, FaceSwap, etc. Each of these methods and tools work over feature matching between the target and the source materials like Skin Colour, eye tracking, lip tracking, etc.

IV. Deepfake Detection

Deepfake’s are detrimental to privacy, hence methods to identify deep fakes were brought to attention as quick as the problem gave rise to fraud. Earlier methodologies used manual labour of stitching frame by frame, but recent advancements in the field of AI has given rise to automation in feature extraction and delivery.

Simple Deepfake detection models started out as a binary classification problem needing the requirement of large datasets containing data of both source and target.But now, you only need to extract features from online sources.For deepfake detection the first step involves face detection which can be achieved by various methods such as Knowledge based methods where rules are formed by the researchers through his/her personal knowledge or feature invariant methods which can be used detect when the orientation pose or angle of the face is varied or template matching method. Template matching method utilizes the edge contours of a basic face shape or appearance-based methods which can be used to classify face and non-face images.

After face detection feature extraction can be carried out using deep learning algorithms which form the base of the deepfake detection program

A. *Deepfake Detection Methods*

Sr. No	**Table 1**			
	Table column subhead	***Authors***	***Description***	***Accuracy***
1	DeepVision: Detection Using Human Eye Blinking Pattern with optical vision. [1]	Yuezun Li, Siwei Lyu (2019, April)	Provides a good insight into how Irregular blinking patterns with optical vision detects Fakes	0.875
2.	Forensics and analysis of deepfake videos.[6]	Jafar, M. T., Ababneh, M., Al-Zoube, M., and Elhassan, A. (2020, April)	Good insights into how latest deepfake methodologies like RCBV and LSTM networks work on	0.91
3.	Detecting deepfakes using Neural ODE and Differential Equation[18]	Steven Fernandes (Jan 2020)	An apt description of proper implementation using Neural ode applied over Videos	0.82
4.	Defending against GAN-based deepfake attacks via transformation-aware adversarial faces [10]	C Yang., Elton Blaird, Lei Ding, Yiran Chen	Gives a good description of how GAN's work and generate Deepfakes and how to counter.	0.86
5.	Face2Face: Real-time Face Capture and Reenactment of RGB Videos[24]	Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt	Uses the Skin colour variations of samples with rgb mix to detect proper fakes.	0.78

B. *Some Additional Methods*

Methods	**Table 2**	
	Key Features	***Data Sets Used***
Using spatial and temporal signatures	CNN model extract feature and do audio embeddings by stacking multiple modules loss function like KullbackLeibler divergence, are used.[a]	FaceForensics+ dataset and 5,600 deepfake audio & videos datasets.
Using audio-visual affective cues	For the face and speech, modality and emotion embedding vectors are extracted and used for deepfake detection.	TIMIT dataset and DFDC
Eye, teach and facial texture	- In this it feat on facial texture differences, eye and teeth area details and missing reflections in deepfake. - Neural network & Logistic regression are used to classify.	A set of morphed videos downloaded from YouTube.
Using phoneme viseme mismatches	- It exploits the discontinuity and inconsistency between the different dynamics and positions of the mouth shape - Focus on sounds associated with the M, B and P phonemes. As deepfakes often incorrectly synthesize it.	Lip-sync deepfakes videos are created using artificial and synthesis techniques, i.e. (A2V) and (T2V)

V. COMPARISON

The below comparison table describes the method that are available to detect deep fake video and its limitation. All the paper that we have referenced so far, follow the same architecture, only thing that makes it unique and efficient from other methods is the feature that has been used. Deep fakes are corrupting the originality of media, as seeing is not what you might believe in anymore. All the method and algorithm has used features such as optical flow [19], face warping artifacts[32], Heart Rate Variations[17] etc. The challenges that are faced by the above-mentioned methods and algorithms are high definition videos, multiple face detection, feature extraction models that take a lot of time to run.

Apart from the fact that the application fields of deep learning extend to other areas of work, it has given rise to such intelligent automation systems that can generate these perpetuated imageries.

Sr. No.	**Table 3**			
	Feature Used	***Purpose***	***Limitations***	***Accuracy***
1.	Human Eye Blinking Pattern[2]	A method to analyze significant changes in eye blinking	Detection can be avoided by purposely incorporating images with closed eyes in training dataset.	0.875
2.	Optical Flow based CNN[19]	Exploits optical flow field dissimilarity and discontinuity as a clue to discriminate between modified videos and real ones	Feature seem to be able to point out some existing dishomogeneous entities between analyzed test cases.	(VGG16) 81.61% (ResNet50) 75.46%
3.	Real-time Face Capture and Reenactment of RGB Videos[24]	The real-time facial reenactment system that requires just monocular RGB input.	Scenes or frames where face is covered with hair and beard are challenging	0.86
4.	Warping Artifacts [32]	Exploiting the face warping artifacts,which are the results of the DeepFake production pipeline.	High resolution video, were difficult to differentiate.	(VGG16) 84.5 (ResNet50) 97.4 using UADFV
5.	Neural ODE [18]	1: Extracting heart rate as a feature 2: Using the above feature to Training Neural-ODE from original videos. 3:Analyzing and Predicting heart rates of the modified videos using previously trained Neural-ODE.	This method has overall accuracy of 75% and require a lot of time to process.	0.91

VI. Architecture

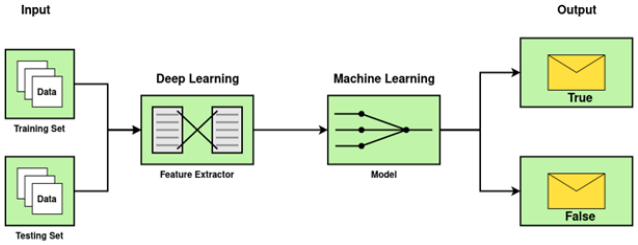

Fig.1 Classification Model

Fig.1 depicts the basics architecture of deepfake detection model.It is divided into three main parts :
1:Dividing the dataset into training and testing sets.
2:Feature extraction using deep learning models.
3: Using features extracted in classification to get outputs.
As the classification model is a binary classification model its outputs will be real or fake.

Input/Datasets: Our dataset includes videos from youtube. It consists of real and fake videos.
Other various datasets available are:
1) DeepFake video dataset DeepfakeTIMIT: This database consist of videos where faces are changed by swapping using the open source GAN-based approach. The dataset consists of two sets of modified videos (64x64) and (128x128) size models each. Among these modified and fake videos, every video has finite subjects around 32 subjects where each subject has 10 more videos of their faces changed or swapped. Every video is 512 × 384 and is roughly 4 secs in length.
2) DeepFake video dataset UADFV: This dataset consists of approximately 100 videos, which have a set of 50 original videos and modified videos . Each video has only one subject and is roughly around 12 secs in length.
Feature Extractor: We are trying to analyze the generation process of a DeepFake video for discovering the anomalies which can be exploited. The program uses a CNN[19] to extract frame-level features which in turn is used to train a RNN for temporal sequence analysis to detect modifications .
Machine learning Model: Model will compare the features extracted from the video frames to the original video frames to train itself to differentiate between original and fake videos. The model is a classification model and will give output as a true (fake video) or false (original video).

VII. Implementation

After referring through a lot of reference paper and decent understanding of deepfake detection as well as generation ,we have made a basic model to test the capabilities of different- different classification models.
In this model we have used a Support vector machine,logistics regression model and multi-Level perceptron[37] and random forest classifier .
The Main Reason for choosing this model are, they all are classification models, have fast training speed, have fast prediction speed, have medium interpretability, medium performance and high performance with the limited datasets.

TABLE 4

Model	***Problem Type***	***Train Speed***	***Predict Speed***	***Interpretability***	***Performance***	***performance with limited data***
SVM	C	slow	moderate	low	moderate	high
LR	C	fast	fast	moderate	lower	high
MLP	Both	slow	moderate	low	high	low
RFC	Both	moderate	moderate	low	moderate	low

C- classification

Both- Classification and Regression.

First of all we have cleaned our datasets using data cleaning techniques to remove missing value and unwanted information to reduce complexity .In order to test and insure that we have a balanced dataset we have looked at the label distribution.As shown in below figure 2.

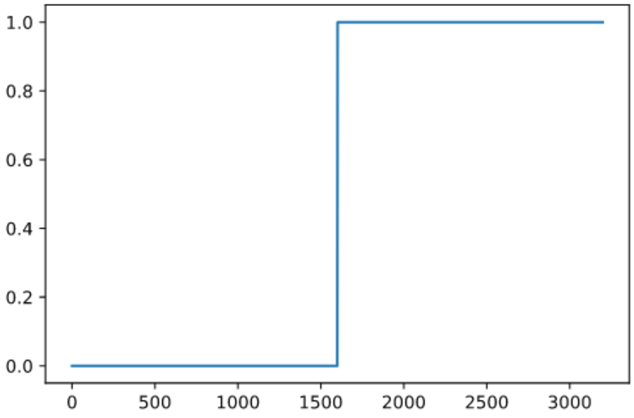

fig - 2 Label Distribution

Than we have divided our dataset into training and testing sets. We have used three-fold cross validation to test different -different classification models.

Logistic: f1=0.787 auc=0.850

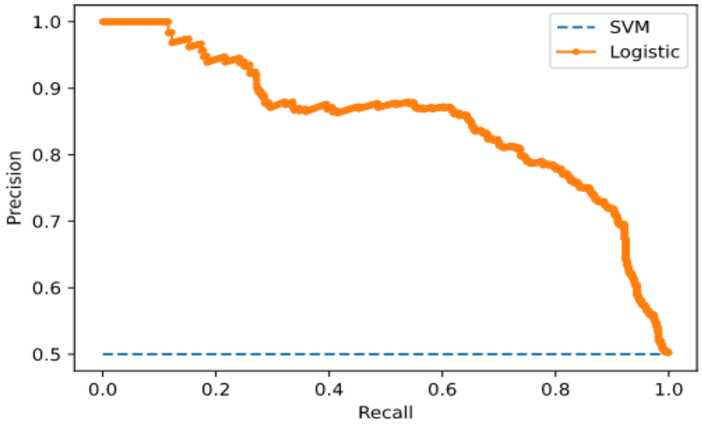

fig - 3 Recall-Precision graph

Test Accuracy per algorithm

Methods	Results	
	Accuracy Achieved	*%*
SVM	0.855	85.5
LR	0.777	77.7
MLP	0.742	74.2
RFC	0.748	74.8

Table 5

VIII. Conclusion

In this work, we have analyzed different methods that are capable of effectively distinguishing between AI-generated morphed videos and original videos. We have evaluated various classification models/methods on several different sets of available DeepFake Videos datasets which demonstrate its efficiency and effectiveness in practice. In the Initial phases we would focus on validation of the model. We have compared various detection models and explored its limitations as well have found the models which transcend the limits.

In this work we have emphasized more on the different methods by which one can detect a deepfake based on the features collected so far. It also gives us a good insight into the best working algorithms that go in hand in hand with the type of dataset used. For a robust algorithm a robust training set is the key.

Based on the observation we have also explained the basic architecture that almost every model has followed. With the continuous evolution in deepfake technology, every new algorithm gives rise to a new branch of loopholes that need to be traced and neutralized iteratively.

Over the years Deepfakes' quality have been evolving rapidly thereby the performance of detection methods need to advance accordingly. We believe that by creating deepfakes perpetrators are not only defaming people but are also defaming AI. Our inspiration is to stop people from defaming AI by spreading awareness and making AI based detection models to prove technology is for the betterment of humankind. We believe that our study will act as a base and derive great insights for the future aspirants under security and forensics field who want to do research in deepfake generation and detection.

References

[1] DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern: https://ieeexplore.ieee.org/document/9072088

[2] Lyu, S. (2018, August 29). Detecting deepfake videos in the blink of an eye. Available at http://theconversation.com/detecting-deepfake-videos-in-the-blink-of-an-eye-101072

[3] DeepFake tf:Deepfake based on Tensorflow.Available At https://github.com/StromWine/DeepFaketf

[4] Park, T., Liu, M. Y., Wang, T. C., and Zhu, J. Y. (2019). Semantic Image synthesis with spatially-adaptive normalization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

[5] DeepFaceLab:Explained And Usage Tutorial.Available at: https://mrdeepfakes.com/forums/thread-deepfacelab-explained-and-usage-tutorial.

[6] Jafar, M. T., Ababneh, M., Al-Zoube, M., and Elhassan, A. (2020,April) Forensics and analysis of deepfake videos. In The 11th International Conference on Information and Communication Systems (ICICS)(pp. 053-058). IEEE.

[7] schroepfer, M. (2019, September 5). Creating a data set and a challenge for deepfakes. Available at https://ai.facebook.com/blog/deepfake-detectionchallenge

[8] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., andNiener, M. (2018). FaceForensics: A large-scale video dataset for forgery detection in human faces.

[9] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., and Choo, J. (2018). StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition

[10] lang, C., Ding, L., Chen, Y., and Li, H. (2020). Defending against GAN-based deepfake attacks via transformation-aware adversarial faces pg 4

[11] Kaliyar, R. K., Goswami, A., and Narang, P. (2020). Deepfake: improving fake news detection using tensor decomposition based deep neural network. Journal of Supercomputing, doi: https://doi.org/10.1007/s11227-020-03294-y.

[12] Tucker, P. (2019, March 31). The newest AI-enabled weapon: Deep-Faking photos of the earth. Available at https://www.defenseone.com/technology/2019/03/next-phaseaideep-faking-whole-world-and-china-ahead/155944/

[13] Fish, T. (2019, April 4). Deep fakes: AI-manipulated media will be weaponised to trick military. Available at https://www.express.co.uk/news/science/1109783/deep-fakesaiartificial-intelligence-photos-video-weaponised-chi

[14] Marr, B. (2019, July 22). The best (and scariest) examples of AI-enabled deepfakes. Available at https://www.forbes.com/sites/bernardmarr/2019/07/22/the-bestandscariest-examples-of-ai-enabled-deepfakes/

[15] Zakharov, E., Shysheya, A., Burkov, E., and Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. arXiv preprint arXiv:1905.08233. Damiani, J. (2019, September 3). A voice deepfake was used to scam a CEO out of $243,000. Available at https://www.forbes.com/sites/jessedamiani/2019/09/03/avoicedeepfake-was-used-to-scam-a-ceo-out-of-243000/

[16] Lyu, S. (2018, August 29). Detecting deepfake videos in the blink of an eye. Available at http://theconversation.com/detecting-deepfake-videosin-the-blink-of-an-eye-101072

[17] Sakthi Kumar Arul Prakash and Conrad S Tucker. Bounded kalman filter method for motion-robust, non-contact heart rate estimation. Biomedical optics express, 2018

[18] Neural Ordinary Differential Equations for Intervention Modeling - https://arxiv.org/abs/2010.08304

[19] Deepfake Video Detection through Optical Flow based CNN - https://openaccess.thecvf.com/content_ICCVW_2019/papers/HBU/Amerini_Deepfake_Video_Detection_through_Optical_Flow_Based_CNN_ICCVW_2019_paper.pdf

[20] Umur Aybars Ciftci and Ilke Demir. Fakecatcher: Detection of synthetic portrait videos using biological signals. CoRR, abs/1901.02212, 2019.

[21] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. (2020). The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397

[22] Gandhi, A., and Jain, S. (2020). Adversarial perturbations fool deepfake detectors. arXiv preprint arXiv:2003.10596.

[23] Hasan, H. R., and Salah, K. (2019). Combating deepfake videos using blockchain and smart contracts

[24] Neekhara, P., Hussain, S., Jere, M., Koushanfar, F., and McAuley, J. (2020). Adversarial deepfakes: evaluating vulnerability of deepfake detectors to adversarial examples. arXiv preprint arXiv:2002.12749

[25] P.-L. Hsieh, C. Ma, J. Yu, and H. Li. Unconstrained realtime facial performance capture. In Computer Vision and Pattern Recognition (CVPR), 2015.

[26] Marissa Koopman, Andrea Macarulla Rodriguez, and Zeno Geradts. Detection of deepfake video manipulation. 08 2018.

[27] Punnappurath, A., and Brown, M. S. (2019). Learning raw image reconstruction-aware deep image compressors. IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI: 10.1109/TPAMI.2019.2903062.

[28]] Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. (2019). Energy compaction-based image compression using convolutional autoencoder. IEEE Transactions on Multimedia. DOI: 10.1109/TMM.2019.2938345

[29]] Chorowski, J., Weiss, R. J., Bengio, S., and Oord, A. V. D. (2019). Unsupervised speech representation learning using wavenet autoencoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 27(12), pp. 2041-2053.

[30]] Faceswap: Deepfakes software for all. Available at https://github.com/deepfakes/faceswap

[31] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 815-823).

[32] Exposing DeepFake Videos By Detecting Face Warping Artifacts -Yuezun Li, Siwei Lyu - https://arxiv.org/abs/1811.00656

[33] Korshunov, P., and Marcel, S. (2018, September). Speaker inconsistency detection in tampered video. In 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 2375-2379). IEEE

[34] Galbally, J., and Marcel, S. (2014, August). Face anti-spoofing based on general image quality assessment. In 2014 22nd International Conference on Pattern Recognition (pp. 1173-1178). IEEE

[35] Zhang, Y., Zheng, L., and Thing, V. L. (2017, August). Automated face swapping and its detection. In 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP) (pp. 15-19). IEEE.

[36] Bai, S. (2017). Growing random forest on deep convolutional neural networks for scene categorization. Expert Systems with Applications, 71, 279-287.

[37] Multilayer perceptron and neural networks https://www.researchgate.net/publication/228340819_Multilayer_perceptron_and_neural_networks