IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

Deepfake Detection through Deep Learning

Prof. Chavan Sir[1], Jayesh S. Pathade[2], Prajwal D. Khandge[3], Nayan N. Khairnar[4], Yuvraj N. Kamble[5]
Professor, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2,3,4,5]
SKN Sinhgad Institute of Technology and Science, Kusgaon (BK), Pune, Maharashtra, India

Abstract: *Deepfakes allow for the automatic generation and creation of (fake) video content, e.g. through generative adversarial networks. Deepfake technology is a controversial technology with many wide reaching issues impacting society, e.g. election biasing. Much research has been devoted to developing detection methods to reduce the potential negative impact of deepfakes. The results suggest that while deepfakes are a significant threat to our society, political system and business, they can be combatted via legislation and regulation, corporate policies and voluntary action, education and training, as well as the development of technology for deepfake detection, content authentication, and deepfake prevention. The study provides a comprehensive review of deepfakes and provides cyber security and AI entrepreneurs with business opportunities in fighting against media forgeries and fake news. Generations and the results were realistic. Moreover, we implemented a DeepFake Detector XceptionNet with minor modifications which achieved 95% accuracy on detecting DeepFakes. At last, we implemented a newly introduced technique in which the DeepFake generation is perturbed through which it can easily fool the deepfake detector..*

Keywords: DeepFakes, Image Animation, DeepFakes Generations, Detection of DeepFakes, GANS, Adversarial Attacks, Fooling DeepFake Detectors

I. INTRODUCTION

In recent years, fake news has become an issue that is a threat to public discourse, human society, and democracy. In the era of fake news, people and society more generally are concerned that they can no longer believe what they see online. In this context, Facebook and Instagram announced a new policy in January 2020 banning AI-manipulated "deepfake" distinguish between real and fake videos. This is the focus of this paper. Deepfakes are synthetic media in which a person in an existing image or video is replaced with someone else's likeness . Figures 1 and 2 show screenshots of some famous deepfake works including Barack Obama calling Trump a "dipshit" and Nicolas Cage's face being swapped to roles in movies in which he did not appear. These videos drew millions of views on Youtube as people were amazed and concerned about how real they appeared. The fast evolution of deepfakes has made both the academic field and technology industry put considerable focus on automated detection of deepfake videos, since more and more people are using deepfakes to generate maany forms of fake information from fakes news to hoaxing of content, e.g. celebrity pornography. The result of the model's evaluations demonstrates a high degree of accuracy in distinguishing real and fake videos however this accuracy is also highly sensitive and depends greatly on the deep fake

II. LITERATURE REVIEW

The term deepfakes was coined from the merger of "deep learning" and "fake", referring to the use of state-of-the-art computer vision methods and deep learning techniques to generate fake videos. Fake videos generated using deepfakes consist of two main categories: face-swapping and face reenactment. Face-swapping involves the automatic replacement of a face in a video or image with another face where the identity of the person in the video changes. This original face swapping method can be dated back to a Reddit user post in 2017 [10]. The method used two encoder-decoder pairs, while the two encoders share parameters during training, the training of the decoders are separated. Specifically, for face scenarios, the original face A enters the encoder, then the decoder of B is connected, and the decoding can be done by replacing the face of B with the face of A. This method required a complex training process and consumed a lot of computing resources. This method only worked if there was a large number of target images and video footage to use as training data. This method has however been widely used as it provided the fifirst deepfake

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

method. The practical implementations of DeepFakes is in various fields. A very positive example of implementations of DeepFakes is a video, a campaign for a social cause "Malaria must die"[5]. Former football star, David Beckham is casted in that video ad where he speaks in 9 different languages, even in a female voice. All because of DeepFakes, that even while speaking in a different language, lip sync of David Beckham is perfect. Audio and video syncs totally fine. This example suggests that DeepFake technology can be used in the Film Industry for dubbing purposes, and even portraying characters which were not actually in front of the camera, or even they never existed. Human like characters can be generated with such technology.

However, it has its darker side too. It gained maximum attention when former president of United States of America, Barrack Obama's DeepFake generated video leaked in public. Animating faces with one source images on driving videos of talking heads is now a very common thing, but with advancements in this field, animation of complete body is now being done. To achieve this, Open Pose library is widely used. On the contrary, a lot of work has been done on the of DeepFakes generations. Recently, a research has been done for the detection of DeepFakes detectors. Literature reveals that traditional adversarial attack using Generative Adversarial Networks are also implemented and the results are quite impressive [1][10]These are a combination of two neural networks which sort of compete with each other [1]. The concept is much appreciated by the experts due to its capabilities. These are capable of generating new stuff. It is also considered as the most interesting idea in the field of AI in last ten years.

The ultimate goal of this work was to identify whether a video was real, or whether it had been generated using deepfake technology. As such, the input to the system clearly has to be a video. However, deep learning models use pictures as input, hence conversion of the system (video) input to the model input needs to be done. This is done through a pre-processing module as shown in Figure 3. In addition to the transformation of input data types, the pre-processing module also needs to take into account the impact that other factors in the video may have on model training. Thus, each frame in the video does not contain just a face. Indeed, the body parts of the person and the background area of the image comprise most of the video frame. These irrelevant features can negatively impact the training of the model. The face area in the image is the focus, and the pre-processing module needs to capture the face in the image as model input. The pre-processing module itself consists of three steps as shown in Figure 4: intercepting frames from the video, detecting faces from these individual frames, and saving face areas as images. Each of these steps is discussed below.

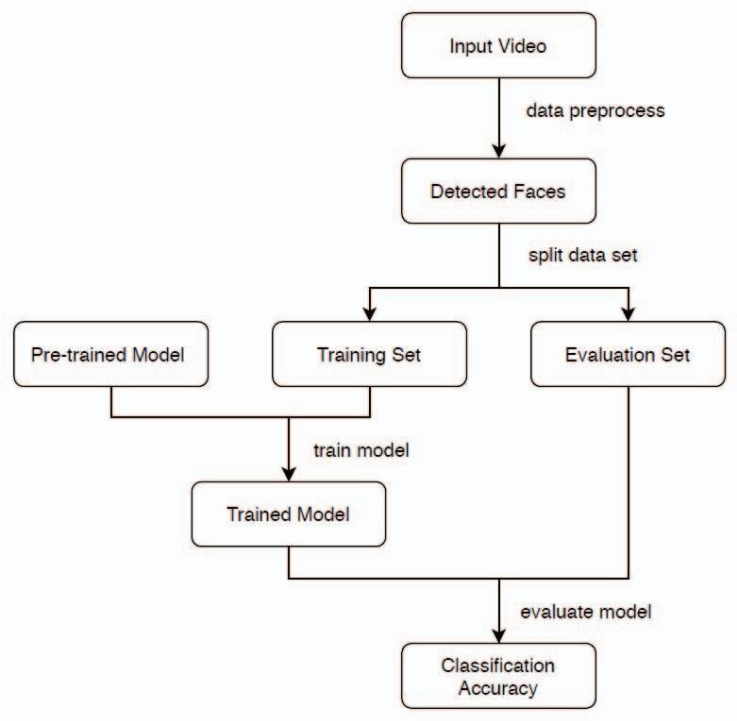

Figure 1: Workflow

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

The first step was to capture the input video into frames. The video capture function provided by the OpenCV Python package was used for this purpose. Since the method explored in this project was based on using a single image as input, inter-frame information was not needed. Besides, the similarity between two adjacent frames is too high hence putting them all into a training set will not only reduce the training effifficiency but also potentially give rise to issues such as overfifitting. The videos selected for this work were all at a frame rate of 30 frames per second. After testing it was identifified that selection of one image from every four frames provided a reasonable approach. The second step was to detect the faces that appear in the image and label them.

The cascade classififier provided by OpenCV was used for this purpose. After testing a variety of classififiers, the haarcascade_frontalface_alt classififierwasThere are many deep learning models and frameworks that are now available. Xception and MobileNet were chosen as models for the experiments in this paper for the following key reasons. Firstly, Xception has a high performance based on its benchmarking performance on the Face Forensics test environment. FaceForensics provides a test environment for researchers to test their trained models. The performance of models trained by different teams are displayed in Figure 6 along with methods they chose. Among these methods, Xception shown a relatively good performance over four different datasets and more importantly, it is open source with extensive documentation, making it easier for model training and tuning. MobileNet was chosen as it shares a similar structure to Xception. They are both based on convolutional neural networks (CNNs) and both employ depthwise and pointwise convolutional layers. The difference between them is that MobileNets has fewer features to increase the effifficiency of the model. lexicons to label each tweet as positive or negative. Using SVM trained from these noisy labeled data, they obtained 81.3% in sentiment classification occuracy.

1)As for technology, deepfakes are the product of Generative Adversarial Networks (GANs), namely two artificial neural networks working together to createsmart assistants (PCM09) and digital doubles of people.

This helps to develop better human relationships and interaction online (CBS03; FRB02). Similarly, the technology can have positive uses in the social and medical fields. Deepfakes can help people deal with the loss of loved ones by digitally bringing a deceased friend “back to life”, and thereby potentially aiding a grieving loved one to say goodbye to her (FOX05; PCM10). Further, it can digitally recreate an amputee’s limb or allow transgender people to better see themselves as a preferred gender (USAT04). Deepfake technology can even help people with Alzheimer's interact with a younger face they may remember (FOX05). Scientists are also exploring the use of GANs to detect abnormalities in X-rays (CNET04) and their potential in creating virtual chemical molecules to speed up materials science and medical discoveries (GRD03).

2) Real-looking media (CNN03). These two networks called ‘the generator’ and ‘the discriminator’ are trained on the same dataset of images, videos, or sounds (GRD03). The first then tries to create new samples that are good enough to trick the second network, which works to determine whether the new media it sees is real (FBR07). That way, they drive each other to improve (PCM05). A GAN can look at thousands of photos of a person, and produce a new portrait that approximates those photos without being an exact copy of any one of them (GRD07). In the near future, GANs will be trained on less information and be able to swap heads, whole bodies, and voices (GRD08; USAT01). Although deepfakes usually require a large number of images to create a realistic forgery, researchers have already developed a technique to generate a fake video by feeding it only one photo such as a selfie (CBS03; CNET07)

III. RESULT AND DISCUSSION

Total of eight deepfake video classifification models were trained, evaluated and compared based on four fake video generation methods and two state of the art neural networks. Each model exhibited satisfactory classifification performance over the corresponding dataset used to train it. Specififically, for the Xception models, the overall fake detection accuracy was above 90% and the model performed slightly better at detecting real videos compared to fake videos with a 2-3% increase in accuracy. The NeuralTextures model was an outlier in this test with the true negative rates. For the MobileNets model the overall detection accuracy was above 90% for videos based on the Deepfakes, Face2Face and FaceSwap platforms but only 88% for videos produced on the NeuralTextures platform.

The models have a similar detection accuracy for fake and real videos. The model trained with NeuralTextures was again an outlier with a 91% true positive rate and 86% true negative rate. A voting mechanism was implemented to utilize the four models together to detect all four types of videos generated by the four mainstream fake video

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

generating methods. This followed a simple model whereby any video classified as fake by any method was ultimately classified as fake. Other models are also possible, e.g. based on ranking and consensus, but given the sensitivity of the models to the specific platforms as shown in Figure 15, this approach was the most sensible to adopt.the benefits and threats ofdeepfake technology, examples of current deepfakes, and how to combat them. In so doing, the study found that deepfakes are hyper-realistic videos digitally manipulated to depict people saying and doing things that never happened. Deepfakes are created using AI, that is, Generative Adversarial Networks (GANs) that pit discriminative and generative algorithms against one another to fine-tune performance with every repetition, and thereby produce a convincing fake (Fletcher, 2018; Spivak, 2019). These fakes of real people are often highly viral and tend to spread quickly through social media platforms, thus making them an efficient tool for disinformation.

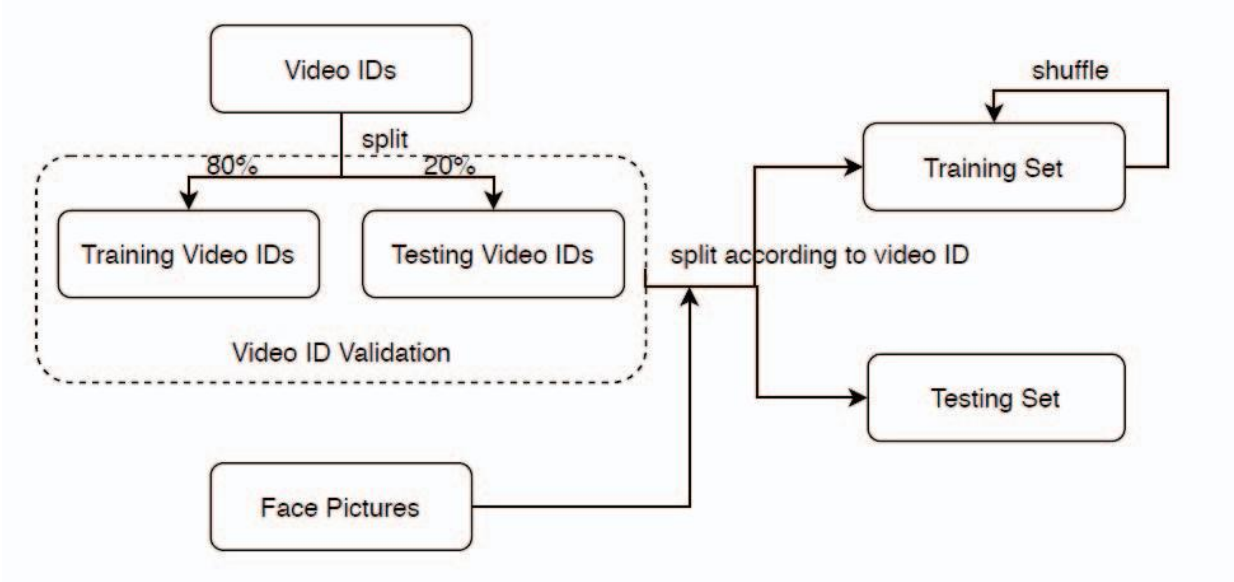

Figure 3: DataPre-Processing

When pre-processing videos, the path of output fifiles was in the form 'pictures/fake/Deepfakes/00132- 00121.png' where 00121 was the picture ID. This implies that this picture was the 121st picture from video 00132. 00132 is the video ID previously, this list was split into two parts with 80% used for training and 20% used for testing.

IV. CONCLUSION

In future work it would be worth exploring the impact of different loss functions and different optimizers on the results. Some researchers have also explored specifific facial features as the dataset to feed in the model, e.g. the eyes, nose, ears or mouth. It would be interesting to compare model performance between a model trained with a whole face and models trained with partial facial features. DeepFakes technique is used inseveral fields. Despite of having its positive implementations in the field of Film Industry, maximizing the production with minimum input, it is still a very dangerous technique and a threat to the world. Manipulative content has been, and still being created with this technology. It is now very easier to defame any famous person by face swapping, or even whole body can be swapped.

V. ACKNOWLEDGEMENT

The authors would like to acknowledge the support and guidance provided by management and guides of SKN Sinhgad Institute of Technology and Science, Lonavala for providing the necessary support and guidance in carrying out this work.

REFERENCES

[1]. Deepfake Detection through Deep Learning,Deng Pan, Lixian Sun, Rui Wang, Xingjian Zhang, Richard O. Sinnott.2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies https://www.researchgate.net/publication/348033118_Deepfake_Detection_through_Deep_Learning

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

[2]. Justus Thies, Michael Zollh¨ofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, "Face2face: Real-timface capture and reenactment of rgb videos," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2387–2395

[3]. A. Bansal, S. Ma, D. Ramanan and Y. Sheikh, "Recycle-GAN: Unsupervised Video Retargeting", Computer Vision – ECCV 2018, pp. 122-138, 2018. Available: https://arxiv.org/pdf/1808.05174.pdf. [Accessed 10 July 2020].

[4]. Xunyu Pan, Xing Zhang, and Siwei Lyu, "Exposing image splicing with inconsistent local noise variances," in 2012 IEEE International Conference on Computational Photography, 2012,pp.110.

[5]. P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, 2017, pp. 1831-1839: IEEE.

[6]. Blinking," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong,2018, pp. 1-7: IEEE.

[7]. 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), Brighton, United Kingdom, 2019, pp. 8261-8265: IEEE.

[8]. H. Li, B. Li, S. Tan, and J. Huang, "Detection of Deep Network Generated Images Using Disparities in Color Components," arXiv preprint :07276, 2018.

[9]. S. McCloskey and M. Albright, "Detecting GAN-generated Imagery using Color Cues,"arXiv preprint rXiv:08247, 2018.