ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse62922020.pdf
https://doi.org/10.30534/ijatcse/2020/62922020

Deepfake Video Detection Using Convolutional Neural Network

Aarti Karandikar, Vedita Deshpande, Sanjana Singh, Sayali Nagbhidkar, Saurabh Agrawal
Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India
karandikara@rknec.edu

ABSTRACT

With the advent of new technological enhancements in artificial intelligence, new sophisticated AI techniques are used to create fake videos. Such videos can pose a great threat to the society in various social and political ways and can be used for malicious purposes. These fake videos are called deepfakes. Deepfakes refer to manipulated videos, or other digital representations produced by sophisticated artificial intelligence, that yield fabricated images and sounds that appear to be real. A deep-learning system can produce a persuasive counterfeit by studying photographs and videos of a target person from multiple angles, and then mimicking its behaviour and speech patterns.

Detecting these videos is a massive problem because of the increasing developments in more realistic deepfake creation technologies emerging every now and then.

The paper aims to solve this problem by proposing a model that analyses the frames of the videos using deep learning approach to detect inconsistencies in facial features, compression rate and discrepancies introduced in the videos while creating them. The model uses a convolutional neural network along with transfer learning to train the model that can catch these instilled errors in the deepfakes. The neural network is trained on these discrepancies induced during deepfake creation around the face. It uses a dataset called “Celeb-DF: A New Dataset for DeepFake Forensics” to train the model. The paper further discusses methods that can be used, in detail, to improve learning by this model.

Key words: Convolutional neural networks, deepfake, digital media forensics, face forensics.

1. INTRODUCTION

Forging of data is nothing new in this era having a backbone made up of artificial intelligence and machine learning. Distortion of reality is becoming a huge problem these days as more and more fake images and videos are emerging everyday on the internet.These deepfakes are getting better with time, to the extent that they cannot be distinguished as fake or real by the human eye, hence are increasingly resistant to detection. While deep-fake technology will bring with it certain benefits, it also will introduce many harms. In an era brimming with so much truth decay, nothing is more dangerous than people taking such videos at their face value. These deepfakes can be used for various malicious purposes such as defamation of celebrities, creating political bias, personal sabotage, intimidation and exploitation, false propaganda, piracy and other vengeful activities. The most targeted feature in a deepfake is the face. Thus, many algorithms and techniques can be used to identify manipulation of faces. These face manipulations can be of two types- Expressions and Identity. In the first type, the expressions of one person are transferred to another in real time. In identity manipulation, faces of two people are swapped. This type of manipulation can be used to spread false information among public by swapping with the face of a famous person.

In this paper, techniques are discussed to detect deepfakes. This is implemented using neural networks such as Convolutional Neural Network (CNN) and image pre-processing techniques. At the end, the result is obtained which distinguishes the real image from the fake one. This is achieved by training the model on a dataset and using a suitable CNN to classify the images as real and fake.

The algorithm works in the following three steps-

- A large dataset of both real and fake videos which are then converted to frames.
- Collection of facial features after alignment and extraction, on which hidden feature extraction is done by a pre-trained Convolutional Neural Network (VGG16 model [1] or OxfordNet).
- A model that trains on the processed dataset and classifies them as real and fake after post processing for videos.

2. DEEPFAKE CREATION

There are numerous ways in which deepfakes can be created. The internet is flooded by softwares to create deepfakes that can be used by users with various technical skills ranging from novice to professional. It is so easy to create deepfakes that any layman can create them, having zero knowledge on the technicalities of the subject. The videos created may be very basic or elementary that can be recognised as fake by just a glance, to highly manipulated videos that cannot be detected even by a keen observer.

These deepfakes are created using artificial intelligence and deep learning methods. They rely on a type of convolutional neural network called auto-encoder which is used for encoding the input image by applying dimension reduction and image compression, and a decoder which reconstructs the image from the constructed representation by the encoder. The auto encoder is a self-supervising algorithm as it uses targets

provided by itself to train on. An upgrade to this method is GAN, i.e. Generative Adversarial Network, an unsupervised deep learning algorithm, which further improves the quality of deepfake created.

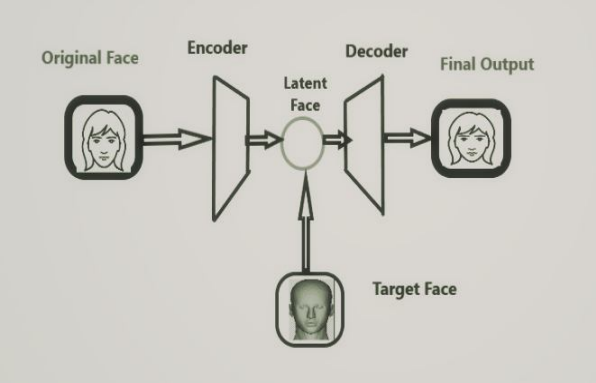

Figure 1: Deepfake Generation using auto encoder-decoder

A GAN is made up of two competing neural networks- generator and discriminator. The generator creates images as close to the real images as possible. The discriminator is then fed a training set containing both real and generated images and it tries to distinguish between them. As it continues to train, the generator makes images which are wrongly classified by the discriminator and the latter gets better at discriminating these images. So they form a pair that learn from each other and improve over time. In this way better quality deepfakes can be created. Deep Convolutional GANs (DCGANs) are even more effective as it uses convolutional layers to increase its efficiency.

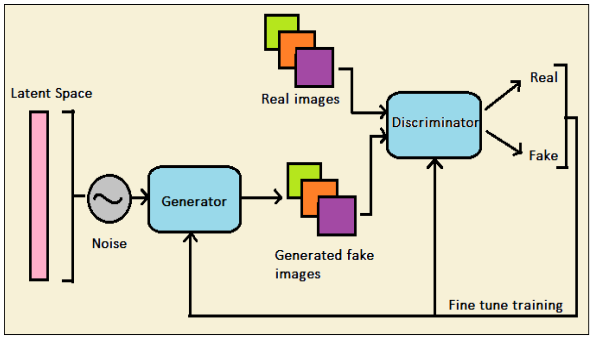

Figure 2: Deepfake Generation using GAN

An existing image of a person can be replaced with someone else's face by superimposing the latter onto the given image using artificial neural networks. There is also a method called FaceSwap which can be used to swap faces in manipulated images and videos. It uses image compression to adjust the superimposed image onto the given image. The colour of the two faces is also matched. The emerging artificial intelligence technologies can even replace expressions of one person to another in real time.

The first application in the direction to create deepfake was FakeApp which allowed users to swap faces with other person. More such applications have been built over time such as FaceSwap, DeepFaceLab, DFaker, FaceSwap-GAN, DeepFake-tf, and many more.

3. RELATED WORK

This section aims to discuss and analyse various techniques that have been used for deepfake detection. Various attempts have been made to detect deepfakes which use deep learning at their core. These approaches work either on detecting faults in video or in separate frames of the video. Approaches which involve image analysis target various parameters like face warping artifacts [4], eye blinking rate [3] and head movements [6]. In 2018, "MesoNet" [5] was developed which used Inception model[11] to detect faults at mesoscopic level. Convolutional Neural Networks (CNN) have thus shown excellent feature extraction properties that can be used by a model to detect deepfake videos.

Various other approaches have used CNN along with other learning models like Recurrent Neural Network (RNN), Long Short-Term Memory Networks (LSTM) and Capsule Network[7] to further improve the accuracy by detecting temporal discrepancies and have shown good results on dataset containing videos generated by FaceSwap and deepfake.

Although various improvements have been made, more robust models are needed to detect deepfakes of lower quality and maintaining considerable accuracy remains a challenge with constantly improving methods in deepfake creation.

This paper aims to discuss successful feature extraction and processing required to detect discrepancies in the deepfake videos.

4. DEEPFAKE DETECTION

DeepFakes use specialized technique which generally modifies fixed areas on face which has to be used as a base for superimposition. The algorithm works in similar way for generating different deepfakes thus leaving some discrepancies during the editing process. Factors like compression changes, lighting differences along with temporal discrepancies like lip and eye movements [3] can be specifically targeted to train models to detect DeepFake videos. Among the methods that have been suggested for DeepFake detection, Convolution Neural Networks (CNN) has been a popular choice. CNNs have shown great ability and scalability for applications regarding image and video processes when compared with other methods for supervised learning in Artificial Intelligence. CNN has the special ability to extract features from an image which can then be used for several applications. Along with feature extraction by a convolutional neural network other supervised learning tools can then be used for final classification for DeepFake to generate better and more precise models for DeepFake Detection.

For practical purposes, transfer learning can be implemented for detecting DeepFakes. Transfer learning

uses pre-trained weights of the neural network for training a fine tuned version of same on different dataset for some specific application. Various pre-trained models have been made open source like VGG Net, Xception, Inception and ResNet. A fine-tuned model trained on a pre-trained convolution network like VGG Net is proposed which is specifically used for Image Analysis on human faces.

Various resources like keras for python provide easy functionality for implementing neural networks for transfer learning. The model uses the pre-processed frame (image) set from the original dataset and implements transfer learning on a fine tuned VGG Net model for detection of DeepFakes.

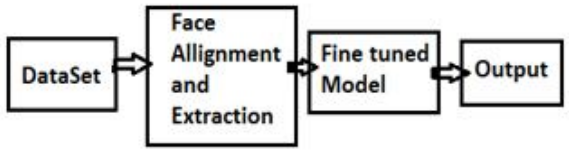

Figure 3: Process flow for deepfake detection

4.1. Dataset

The Celeb-DF dataset consists of 408 real and 795 synthesized videos that are made using modified DeepFake generation algorithm. The average length of the videos is 13 seconds with 30fps frame rate. The synthesized videos consist of low visual artifacts so that the quality of videos is greater than the previous datasets, where the videos are of high resolution and high number of visual artifacts that makes it easier to detect deepfakes. This dataset consists of lower quality deepfakes that makes the problem more challenging.

Table I: Contents of Dataset

1.Synthesized Videos	795
2.Real Videos used for creating deepfakes	158
3.Real Videos from Youtube	250

Figure 4: Generated fake samples from Celeb-DF dataset

Figure 5: Generated real samples from Celeb-DF dataset

4.2. Proposed Method

This paper proposes a method to train the classifier based on video frames as input. The frames are passed through face extraction and alignment fragment and then passed to the classifier for training.

The dataset is pre-processed before training the model. This involves face alignment and extraction. The proposed model targets faults induced during deepfake creation around the face outline. Thus face extraction will extract the area that needs to be processed. Face alignment is used to account for different head positions that the target person may have in the deepfake video.

The paper proposes a fine tuned convolutional model to be trained of the dataset after preprocessing.

The proposed classifier consists of VGG-16 model as its base which is then appended with batch normalization, dropout and a custom two node dense layer. The two nodes in last dense layer in the architecture proposed are used for two final classes (real and fake). Batch normalization layer is used for normalization and scaling for inputs from previous layer. Further, dropout is added to reduce overfitting and better optimization of weights. The dropout layer will randomly send some nodes as off from the previous layer at every epoch. This will lead to better training as some randomness is induced by this layer while updation of weights. Also, Adam Optimizer is used as it gives best learning for this scenario when compared with other optimizers like Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square (RMSProp). It has benefits of RMSProp along with added momentum parameter. Learning rate should be kept optimum (around 0.001) for successful feature extraction and training.

Transfer learning is then implemented on this model for a pre-processed dataset. The image is passed through neural network that assesses the image and extracts simple features from it. These features are then checked for anomalies at the pixel level that are introduced while creating the fake image like lossy compression scheme, artifacts introduced during image warping and subtle colour changes. These areas of discrepancies can be highlighted by the use of dlib library.

After training, the results for all the frames for a training video are used for post processing where video analysis is done.

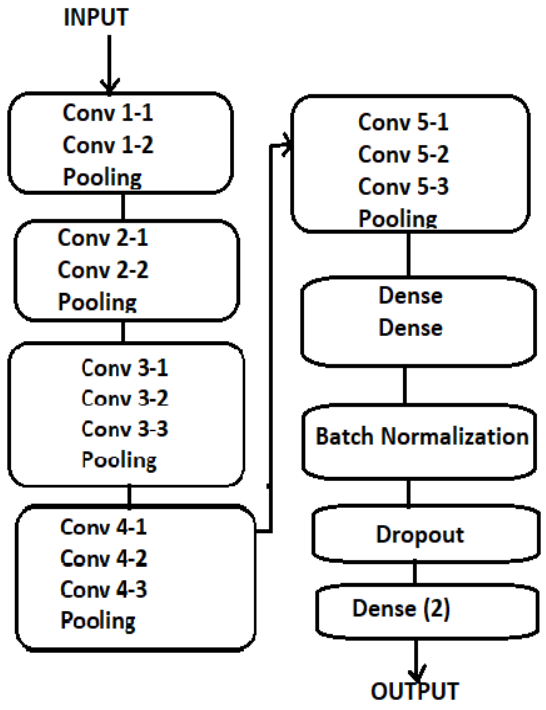

Figure 6: Architecture of the fine-tuned VGG model used

4.3. Fake Video Detection

The model tries to target DeepFake videos by converting them to a fixed number of frames. The Fake Image detection model can now be used for individual frames of the video. Information gained from each individual frame can be aggregated to get a combined conclusion about the class for the video. The weights in the network are thus updated based on this collective information. This method is simpler to implement when compared with video processing in neural network itself. The parameter indicating the number of frames to be used can be changed by observing the learning process of the model.

For any test video, the video can be processed frame by frame and the predictions over each frame can be used to give final classification. Processing videos leads to increased accuracy as different versions of similar input are now part of the dataset which is fed to the neural network. Also, different image transformation operations like zooming, flipping and rotating the frame with small angle are also considered to create a richer dataset as the output class for the frame remains same after applying these transformations.

Coupling the feature extraction and pre-processing model along with models to detect temporal features like recurrent neural network (RNN) can lead to better accuracy.

5.RESULT ANALYSIS

The model described in the paper results in accuracy of around 70% based on the features learned by image analysis. Thus features learned at this step can be used for further temporal analysis to effectively detect deepfakes.

The model was trained on around 4000 real frames and 5000 fake frames. The test data constituted 30% of the available data.

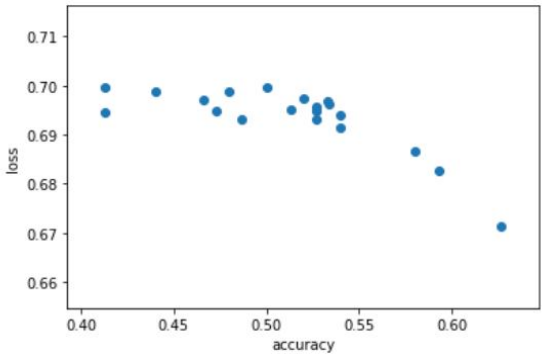

Figure 7: Loss vs. Accuracy curve for 20 epochs

Figure 7 shows the plot of categorical cross entropy loss function value against the accuracy for the model while training for 20 epochs. With each epoch, the loss decreases while the accuracy increases. A better trained model can be obtained by increasing the epochs for better learning.

Thus the model shows fairly good accuracy over the dataset containing images with low resolution. Low resolution images make learning difficult and more work needs to be done in building a high resolution dataset for deepfakes.

Other issue appears in the form of compression. Although signature styles for compression in deepfakes can help in training the model, it can still lead to errors in learning and thus has to be dealt with using techniques for temporal analysis.

6. DISCUSSIONS, CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

DeepFake detection is a major need in today's world and needs considerable detection techniques as detecting deepfakes will become more challenging in the future. As deepfakes can have major social and political impact improvements should be made continuously in its detection techniques.

In this paper, proposed method uses transfer learning on VGG-16 model to train the dataset and focus on facial manipulation for detection of forgery. Transfer learning is essential as the model should be trained in considerable amount of time and should require minimum resources to give the desired accuracy for its classification over varied examples in the dataset. The proposed model works well and is able to successfully gather features required for further processing to test for deepfakes. For improving

the performance, further research can be done on detecting temporal and audio discrepancies and then using this combined information with features extracted from image processing module.

It is observed that the accuracy of the proposed model decreases with low quality images and with medium quality videos the accuracy needs to be further increased using combined models for training on temporal parameters. Thus better dataset with improved quality will lead to better training.

Various ensemble learning techniques can also be implemented to further increase the accuracy of the model and account for variance in the dataset. Aggregation of results over each frame and over different learning models will thus give best results.

The authors hope that the presented techniques for analysis on deepfakes will pave the way for further research in the field of image & video forgery and digital media forensics.

REFERENCES

1. Karen Simonyan and Andrew Zisserman, "**Very Deep Convolutional Networks for Large –Scale Image Recognition**", ICLR 2015, arXiv:1409.1556v6 [cs.CV] 10 Apr 2015

2. Chesney, Robert and Citron, Danielle Keats, **Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security** (July 14, 2018). 107 California Law Review (2019, Forthcoming); U of Texas Law, Public Law Research Paper No. 692; U of Maryland Legal Studies Research Paper No. 2018-21

3. Yuezun Li, Ming-Ching Chang and SiweiLyu,In Ictu Oculi: **Exposing AI Generated Fake Face Videos by Detecting Eye Blinking**, arXiv:1806.02877v2 [cs.CV] 11 Jun 2018

4. Yuezun Li, and SiweiLyu, "**Exposing DeepFake Videos By Detecting Face Warping Artifacts**", arXiv:1811.00656v3 [cs.CV] 22 May 2019
https://doi.org/10.1016/S0969-4765(19)30137-7

5. DariusAfchar, Vincent Nozick, Junichi Yamagishi and Isao Echizen, "**MesoNet: a Compact Facial Video Forgery Detection Network**", arXiv:1809.00888v1 [cs.CV] 4 Sep 2018
https://doi.org/10.1109/WIFS.2018.8630761

6. Xin Yang ,Yuezun Li and SiweiLyu, "**Exposing Deep Fakes Using Inconsistent Head Poses**" , ICASSP 2019 - 2019 IEEE ICASSP,17 May 2019

7. Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen, "**Use of a capsule network to detect fake images and videos**", arXiv:1910.12467v2 [cs.CV] 29 Oct 2019

8. FalkoMatern , Christian Riess and Marc Stamminger, "**Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations** ", 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)
https://doi.org/10.1109/WACVW.2019.00020

9. Jessica and Silbey Woodrow Hartzog , "**The Upside of Deep Fakes**", Maryland Law Review, Volume 78 issue 4,2019

10. Schwartz, Oscar (12 November 2018). "**You thought fake news was bad? Deep fakes are where the truth goes to die**". The Guardian.

11. Sik-Ho Tsang, "Review: Inception-v3 — 1st Runner Up (Image Classification) in ILSVRC 2015" ,https://medium.com/@sh.tsang/review-inception-v3-1st-runner-up-image-classification-in-ilsvrc-2015-17915421f77c.

12.PavelKorshunov, Sebastien Marcel, "**DeepFakes: a New Threat to Face Recognition? Assessment and Detection",** citing arXiv: 1812.08685[cs.CV], 20 Dec 2018.

13.David Güera,Edward J. Delp, **"Deepfake Video Detection Using Recurrent Neural Networks",** 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).
https://doi.org/10.1109/AVSS.2018.8639163

14.Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, Cristian Canton Ferrer, "**The Deepfake Detection Challenge (DFDC) Preview Dataset**", arXiv:1910.08854 [cs.CV], 19 Oct 2019.

15. PavelKorshunov and Sebastien Marcel, "**Vulnerability assessment and detection of Deepfake videos**", IAPR International Conference 2019.

16.ThanhThi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, DucThanh Nguyen, SaeidNahavandi, "**Deep Learning for Deepfakes Creation and Detection**", arXiv:1909.11573 [cs.CV], 25 Sep 2019.

17.EkraamSabir, Jiaxin Cheng, AyushJaiswal, WaelAbdAlmageed, IacopoMasi, Prem Natarajan , "**Recurrent Convolutional Strategies for Face Manipulation Detection in Videos**", arXiv:1905.00582 [cs.CV], 2 May 2019**.**

18. Shuo Yuan,Xinguo Yu,Abdul Majid, "**Robust Face Tracking Using Siamese-VGG with Pre-training and Fine-tuning**", 4th International Conference on Control and Robotics Engineering (ICCRE), 20-23 April 2019.
https://doi.org/10.1109/ICCRE.2019.8724212

19.Francesco Marra,Diego Gragnaniello, Davide Cozzolino,Luisa Verdoliva, "**Detection of GAN-Generated Fake Images over Social Networks**", IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 10-12 April 2018.

20.Shubhangi Tirpude1 , Naman Vidyabhanu2 , Hashir Sheikh3 ShoebPathan 4 , Zeeshan Ali syed5 , Shivam Singh, "**Abnormal X-Ray Detection System using Convolution Neural Network**", International Journal of Advanced Trends in Computer Science and Engineering, ISSN 2278-3091, Volume 9, No. 1, January-February 2020.