A report on

Methods for Detecting Deepfake Images

Under the guidance of
Dr. Vrijendra Singh
Associate Professor
Head of Dept. (Information Technology)
IIIT Allahabad

Submitted by
Siddhanta Biswas
MIT2019014
M.Tech (Information Technology)
IIIT Allahabad

I. INTRODUCTION

Over the last few decades, the rise of the internet and with smartphones and tablets and websites like Youtube and social media it's easier than ever to create digital content and publish it online. The digital content shared primarily are digital images and videos. With such popularity of digital media, apps and techniques that perform facial manipulation are readily available that can be used in forgery cases like falsifying identities, revenge porn, etc. Although many techniques are available to detect image forgery which work quite well, video forgery detection is still a difficult task, mostly in case of low quality videos available online.

Detection performance on compressed low quality images are of utmost importance as videos available on the internet are primarily all compressed formats. Even if manipulations are performed on HQ uncompressed videos, when posted online, videos get compressed and artifacts are introduced which make the detection work fairly more difficult. Having made these observations, I feel that it might be possible to significantly increase the performance of the current architectures on highly compressed low quality images.

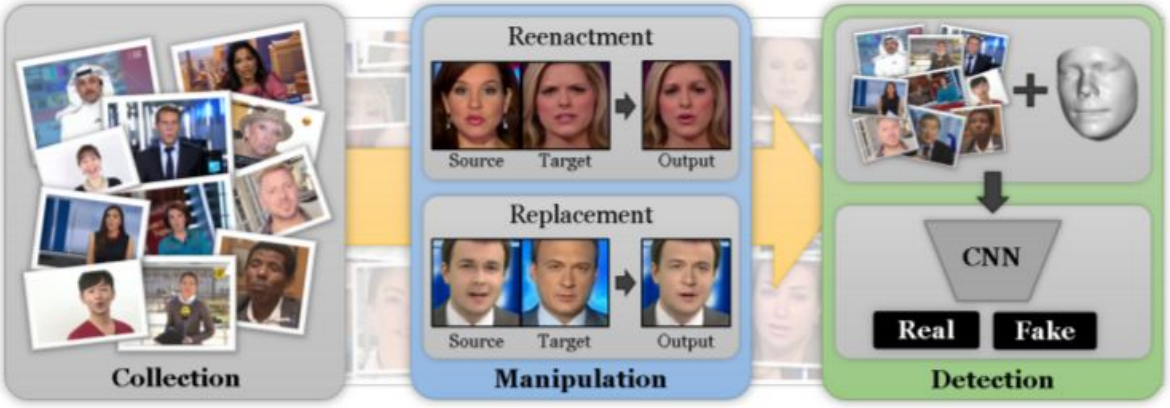

Figure 1: How Deepfakes are generated and detected

II. PROBLEM STATEMENT

Deepfake is a deep learning based technique which aims to replace the face of a targeted person by the face of someone else in digital media. Deepfake detection techniques in recent years have improved largely while still having relatively lower accuracy for low quality images. In this work, I am aiming to significantly increase the accuracy of manipulation detection in low quality images.

III. LITERATURE SURVEY

Various methods have already been developed to detect facial manipulation in videos in the last few years. Facial manipulation can be recognised into these broad categories as found in paper [1]: Entire Face Synthesis, Identity Swap, Attribute Manipulation, Expression Swap.
Among these, my interest of study is methods to detect Identity Swap. Identity Swap consists of replacing the face of one person in a video with the face of another person. Identity Swap can be done by classical computer graphics-based techniques such as FaceSwap and also deep learning techniques known as DeepFakes.

In paper [2], the authors have tweaked the XceptionNet architecture to better suit the objective of detecting manipulated images. This CNN-based system which was tweaked for the face manipulation detection task. In general, the detection system based on XceptionNet architecture provided the best results in both types of manipulation methods, DeepFakes and FaceSwap, with FaceForensics++ dataset. But, the accuracy declined for compressed low quality images significantly. Also, from paper [6], on the CelebDF dataset, this method didn't have any significant edge in performance as compared to some other notable methods like MesoNet.

In paper [3] the authors have proposed MesoNet, which is a CNN-based network to detect face tampering in videos. MesoNet consists of two architectures, Meso-4 and MesoInception-4 which, according to the authors, have achieved the best classification scores among all the tests conducted by them. The MesoNet network has much fewer layers than some other networks and is much simpler and efficient. But according to benchmarks in paper [2], it still suffers from lower accuracy in case of low quality videos.

In paper [4], we see that different methods of image forgery and manipulation leave certain subtle traces in the resultant images in the form of recurrent micro patterns. Similar to steganalysis where hidden messages are extracted by the means of steganography, the authors cast the hand-crafted steganalysis features to a CNN-based network which can detect manipulations based on the "hand-crafted" features that the network is designed to detect explicitly. As seen from benchmarks in paper [1] and [2], this method has low accuracy.

In paper [5], the authors propose a new universal approach for performing image editing detection that is capable of automatically learning traces left by editing. This is done by a new form of convolutional layer designed to suppress an image's content and adaptively learn manipulation detection features. Using this new convolutional layer, this CNN architecture is capable of automatically learning how to detect multiple image manipulations without relying on pre-selected features or models. This method too has the similar trend of having lower accuracy for low quality images.

IV. PROPOSED METHODOLOGY

Detecting facial manipulation in videos can be done in a wide array of methods. Among these, MesoNet and XceptionNet provide the most accuracy while still falling short for videos with higher compression as shown in paper [2] and [6]. The model which I propose to develop should have significantly more accuracy, especially for low quality compressed video formats than the current architectures

Collection of video dataset will be the first objective. The dataset will be containing videos (both real and forged) compressed using H.264 codec with varying lengths and compression rates. For a good enough dataset with sufficient assortments of low quality compressed videos, I've chosen the FaceForensics++ dataset [2], the Deepfake Database dataset by the authors of paper [4] and the Celeb-DF dataset [6]. All the video datasets chosen have H.264 as the compression codec as most of the videos found on the internet are compressed using this codec. The number of videos and if a few other datasets are to be added would be decided later.

Data preprocessing would be done by extracting faces from individual frames in the videos. This can be done with techniques described in paper [3]. Extracted faces must be aligned as well. Cleaning of the extracted dataset, like removing wrongly detected faces or misaligned faces can be done manually.

The processed and cleaned dataset will be used to train the original MesoNet architecture to get a base reading of its accuracy. These readings will serve as the control. The important part of the research comes as making changes in the original MesoNet architecture. Newness and creativity will be implemented by changing the layer architecture and tweaking the hyper-parameters to better extract facial features from compressed lower quality videos. For starters, more convolution layers can be added for extracting more features. In case of overfitting, I'll introduce regularisation techniques such as dropout. Every change made in the model which would lead to better results than previous attempts would be recorded till no more improvement could be made. Tests with HQ images will be conducted parallely to keep an eye if the changes being made do not hurt the already high accuracy for HQ videos negatively.

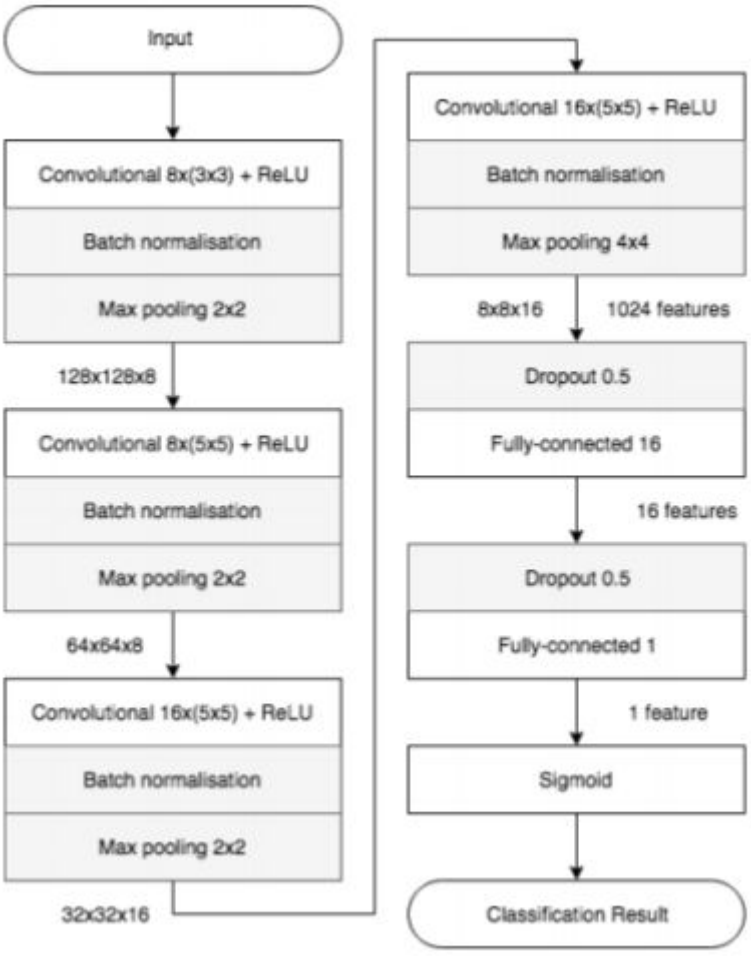

Figure 2: The network architecture of Meso4.

V. EXPERIMENTAL SETUP

The initial objective is to reproduce the results obtained by the authors of paper [3] as closely as possible.To obtain that the same dataset was used, along with the same exact network architecture.

Dataset Used

The dataset used is called the Deepfake Database, created by Darius Afchar et al. , the authors of paper [3]. The dataset has been created from 175 rushes of forged videos that have been collected from different platforms. Their duration ranges from two seconds to three minutes and have a minimum resolution of 854 × 480 pixels. All videos are compressed using the H.264 codec but with different compression levels, which puts us in real conditions of analysis. All the faces have been extracted using the Viola-Jones detector and aligned using a trained neural network for facial landmark detection. The dataset has then been doubled with real face images, also extracted from various internet sources and with the same resolutions. As much as possible, the same distribution of good resolution and poor resolution images were used in both classes to avoid bias in the classification task.

Set	Fake Class	Real Class	Total
Deepfake training	5111	7250	12361
Deepfake testing	2889	4259	7148

Table 1: Cardinality of each class in the Deepfake Database dataset

Classification Setup

The classification task being attempted is a binary classification with two classes, Real and Fake. Classification was attempted with the Meso4 network and was implemented using Python 3.8 and Keras API. Following are the implementation details:

- Input Shape = 200 x 200 x 3
- Testing set was divided into 80:20 ratio for training and model validation purposes simultaneously.
- Epochs = 200
- Batch Size = 75
- Learning Rate = 0.001 (fixed)
- Optimizer: ADAM with default parameters (β1 = 0.9 and β2 = 0.999)

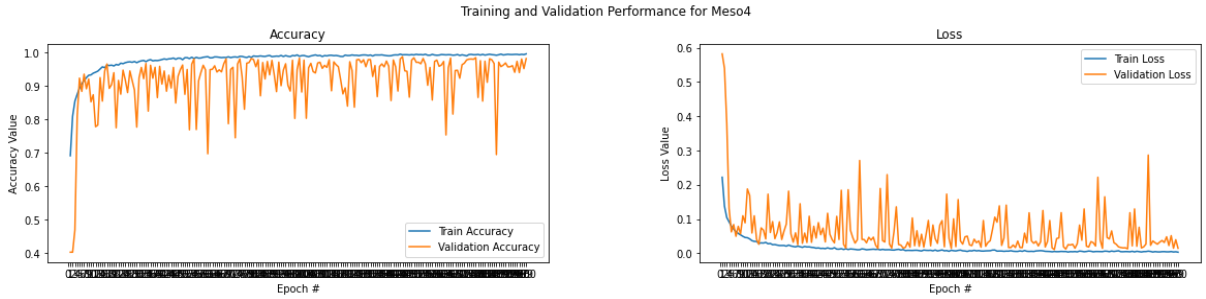

Figure 3: Accuracy and Loss metrics for training and validation

VI. RESULTS AND ANALYSIS

Classification was done on the Testing set. Results of classification are shown in Table 2 and 3. From the classification scores, we can see that albeit being somewhat good scores, my implementation lacks accuracy from the original implementation as shown in the paper [3]. Though the fake images are being detected quite well, the real images are not being detected with good accuracy. This might be due to certain differences in the training phase, like the use of higher resolution images as input, longer training with much larger epochs and also using Learning Rate decay.

Fake Images	96%
Real Images	84%
Total Accuracy	89%

Table 2: Classification scores of my Meso4 implementation on Deepfake dataset

True Positive	2734
False Positive	111
False Negative	664
True Negative	3595

Table 3: Confusion Matrix showing classification scores of Meso4 implementation on deepfake dataset. “Positive” label corresponds to “Fake” class and “Negative” label corresponds to “Real” class

Shown below is the confusion matrix of the prediction results of my Meso4 implementation on the deepfake dataset. Here “Positive” label corresponds to “Fake” class and “Negative” label corresponds to “Real” class. The number of false negatives is particularly high.

Figure 4: Confusion Matrix showing classification scores of Meso4 implementation on Deepfake dataset. “Positive” label corresponds to “Fake” class and “Negative” label corresponds to “Real” class

VII. FUTURE PLAN OF ACTION

From the classification results obtained, a few points are certainly obvious. More faithful reproduction of the original work needs to be done to get comparable results. Comparable results will help me better judge the performance of my model to that of the original benchmarks as shown in paper [3]. After that is taken care of, the following are the immediate plans of action that I propose to work towards my goals:

- Improve present detection rates, especially for Real Images which are being detected with significantly lower accuracy.
- Testing the model on a different dataset (i.e. FaceForensics++) [2] with higher compression levels to get an idea of how this model performs on compressed low quality media.
- Proposed model as shown in Figure 5, consists of adding a bunch of Conv Layers in the third and fourth blocks of the current architecture before Pooling (similar to what VGG-16 does)
- If the number of trainable params increase substantially, resulting in huge training times, 1x1 conv layers may be used to decrease the number of feature maps while still retaining their salient features.

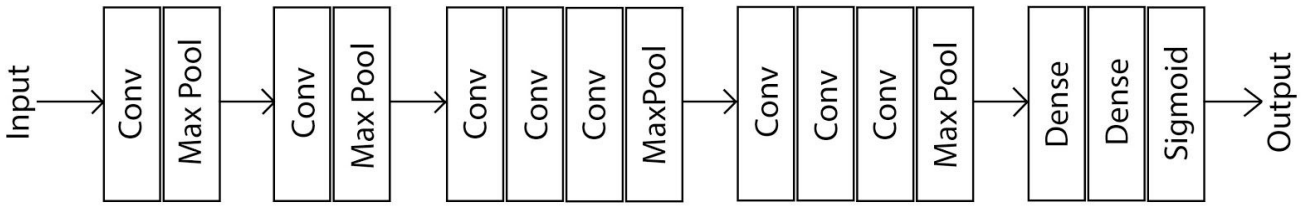

Figure 5: Architecture of my proposed model for deepfake classification, based on the Meso4 network, with several convolutional blocks added (inspired by VGG-16)

REFERENCES

[1] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. arXiv preprint arXiv:2001.00179 (2020)

[2] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and ¨ M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE/CVF International Conference on Computer Vision, 2019

[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in Proc. IEEE International Workshop on Information Forensics and Security, 2018

[4] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting Residual-Based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detection," in Proc. ACM Workshop on Information Hiding and Multimedia Security, 2017.

[5] B. Bayar and M. Stamm, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer," in Proc. ACM Workshop on Information Hiding and Multimedia Security, 2016

[6] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.