MesoNet: Facial Video Counterfeit Detection Network Analysis

Ravi Sankar Gogineni

rgoginen@ttu.edu

Abstract

A method to automatically and efficiently detect face tampering in videos, and particularly focuses on two recent techniques which are used to generate hyper realistic forged/counterfeit videos: Deepfake and Face2Face.The popularization of smartphones and the growth of social networks have made digital images and videos very common digital objects. Tremendous use of digital images has been followed by a rise of techniques to alter image contents, using editing software. The field of digital image forensics research is dedicated to the detection of image forgeries in order to regulate the circulation of such falsified contents. Deep learning performs very well in digital forensics and disrupts traditional signal processing approaches. Addressing the problem of detecting these two video editing processes, Deepfake follows Face2Face. Up to today, there is no other method dedicated to the detection of the Deepfake video falsification technique.

- Deepfake is a technique which aims to replace the face of a targeted person by the face of someone else in a video. The core idea lies in the parallel training of two autoencoders.
- Face2Face is a reenactment method are designed to transfer image facial expression from a source to a target person. Face2Face.

The final image synthesis is rendered by overlaying the target face with a morphed facial blend shape to fit the source facial expression. We propose to detect forged videos of faces by placing our method at a mesoscopic level of analysis. At a higher semantic level, human eye struggles to distinguish forged images, especially when the image depicts a human face. That is why we propose to adopt an intermediate approach using a deep neural network with a small number of layers. The two following architectures have achieved the best classification scores among all our tests, with a low level of representation and a surprisingly low number of parameters:

- Meso-4
- MesoInception-4

Deliverables:

By the end of September: Gathering the requirements and scrutinizing few related papers.

Contribution of each person:

Sai Meghana Akula: Gathering and analysing the related papers on forged video detection using deep learning techniques, theory on Convolution neural networks and recurrent neural networks. Dataset for testing released by Kaggle on deepfake.

Venkata Veera Siva Dasari: Analysis of the Forged images generated by using Deepfake and fake2fake techniques.

Ravi Sankar Gogineni: Meso-net and meso-inception 4 techniques, number of layers to be chosen and conduct the experiment to obtain better results and to find the forged percentage.

Below are the few papers that we inferred to acquire required information to obtain better results:

- Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. [Laith Alzubaidi Jinglan Zhang, Amjad J. Humaidi]
- Deep Learning for Deepfakes Creation and Detection: A Survey [Thanh Thi Nguyena, Quoc Viet Hung Nguyenb, Dung Tien Nguyena, Duc Thanh Nguyena, Thien Huynh-Thec, Saeid Nahavandid, Thanh Tam Nguyene, Quoc-Viet Phamf, Cuong M. Nguyeng
- Basic information that we need to know to carry forward the experiment is about the deep learning techniques (CNN and RNN).

Convolution Neural Networks (CNNs):

CNN is an advanced and high-potential type of the classic artificial neural network model. It is built for tackling higher complexity, pre-processing, and data compilation. It takes reference from the order of arrangement of neurons present in the visual cortex of an animal brain.
The CNNs can be considered as one of the most efficiently flexible models for specializing in image as well as non-image data. These have four different organizations:

- It is made up of a single input layer, which generally is a two-dimensional arrangement of neurons for analysing primary image data, which is similar to that of photo pixels.
- Some CNNs also consist of a single-dimensional output layer of neurons that processes images on their inputs, via the scattered connected convolutional layers.
- The CNNs also have the presence of a third layer known as the sampling layer to limit the number of neurons involved in the corresponding network layers.
- Overall, CNNs have single or multiple connected layers that connect the sampling to output layers.

The CNNs are adequate for tasks, including image recognition, image analyzing, image segmentation, video analysis, and natural language processing.

Recurrent Neural Networks (RNNs):

The RNNs were first designed to help predict sequences, for example, the **Long Short-Term Memory (LSTM)** algorithm is known for its multiple functionalities. Such networks work entirely on data sequences of the variable input length.
The rnn's puts the knowledge gained from its previous state as an input value for the current prediction. Therefore, it can help in achieving short-term memory in a network, leading to the effective management of stock price changes, or other time-based data systems.
LSTMs: Useful in the prediction of data in time sequences, using memory. It has three gates: Input, Output, and Forget.
Gated RNNs: Also useful in data prediction of time sequences via memory. It has two gates— Update and Reset.

Deepfake:

Traditional face tampering is a tedious and time-consuming process, which requires professional video editing tools and professional knowledge. With the continuous development of computer hardware and deep learning, image synthesis has been an enormous breakthrough.

As the advance of generative adversarial networks and Auto-Encoders, these techniques are increasingly being applied to Deepfake, enabling the creation and rapid distribution of high-quality video tampering content. While Deepfake technology has advanced society in some ways, there are concerns about the harmful effects of its misuse. As Deepfake requires only a small number of face photos to enable video face-swapping, some malicious users have taken advantage of the data available on the internet to generate numerous fake videos. Many deceptive face-swapping videos pose a substantial potential threat to national security, social stability, and personal privacy.

If compressed Deepfake videos, it would be difficult for us to detect them effectively. To address the problem, the forensics of compressed Deepfake videos becomes a meaningful and challenging task. Many Deepfake detection methods have been proposed and achieved good detection performance. However, from the perspective of symmetry, the formidable challenges for Deepfake detection still exist when under compression attacks, the detection performance of most detectors dramatically decreases due to the loss of image feature information.

The current detection methods can be divided into two categories according to different feature extraction methods: Deepfake video detection methods based on hand-crafted and Deepfake video detection methods based on deep learning.

These methods are further divided into two categories:

- Frame-level detection methods and
- Video-level detection methods.

Proposing propose Meso-4 and MesoInception-4 networks.

Deepfake detection is performed with the help of image mesoscopic features. The method is trained and tested on the Deepfake dataset constructed by the author and has achieved good detection results.

Face2face:

These methods are designed for the re-enactment of image facial expression from a source to a target person.

Facial re-enactment refers to the modifications brought to the target actions in the form of change of movement of the head, lips, and facial expression.

It performs a photorealistic and marker less facial re-enactment in real-time from a simple RGB-camera.

The program first requires few minutes of pre-recorded videos of the target person for a training sequence to reconstruct its facial model.

the pro- gram tracks both the expressions of the source and target video.

The final image synthesis is rendered by overlaying the target face with a morphed facial blend shape to fit the source facial expression.

This section presents several effective approaches to deal with either Deepfake or Face2Face. It seems like that these two problems cannot be efficiently solved with a unique network. However, thanks to the similar nature of the falsifications, identical network structures for both problems can yield good results.

The below forged video link is the test subject considered to run the project. Detection test will be used on the video link:

https://www.youtube.com/watch?v=VWrhRBb-1Ig&t=98s

Meso – net:

To detect forged videos of faces by placing our method at a mesoscopic level of analysis. Indeed, microscopic analyses based on image noise cannot be applied in a compressed video context where the image noise is strongly degraded. Similarly, at a higher semantic level, human eye struggles to distinguish forged images, especially when the image depicts a human face. That is why we propose to adopt an intermediate approach using a deep neural network with a small number of layers. This network begins with a sequence of four layers of successive convolutions and pooling, and is followed by a dense network with one hidden layer. To improve generalization, the convolutional layers use ReLU activation functions that introduce non-linearity's and Batch Normalization to regularize their output and prevent the vanishing gradient effect, and the fully-connected layers use Dropout to regularize and improve their original quality. In total, there are 27,977 trainable parameters for this network.

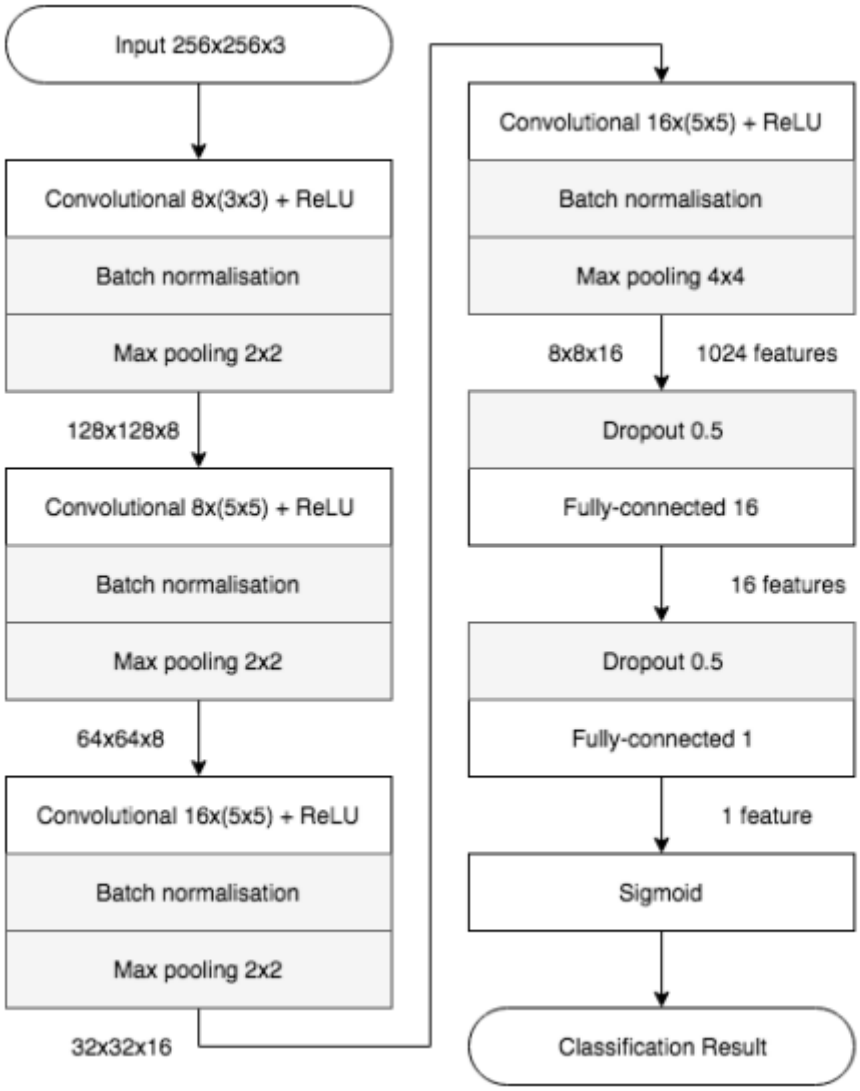

Figure: The network architecture of Meso-4. Layers and parameters are displayed in the boxes, output sizes next to the arrows.

MesoInception-4:

An alternate structure involves using a modified version of the inception module that Szegedy et al. presented to replace the first two convolutional layers of Meso4. The module's goal is to expand the function space in which the model is optimized by stacking the output of many convolutional layers with various kernel shapes. To avoid high semantic, we suggest using 33 dilated convolutions rather than the 55 convolutions of the original module. We have added 11 convolutions before dilated convolutions for dimension reduction and an additional 11 convolutions in parallel that acts as skip-connection between succeeding modules. This method of

using dilated convolutions with the inception module can be found in as a way to deal with multi-scale information. In Figure, more information is shown.

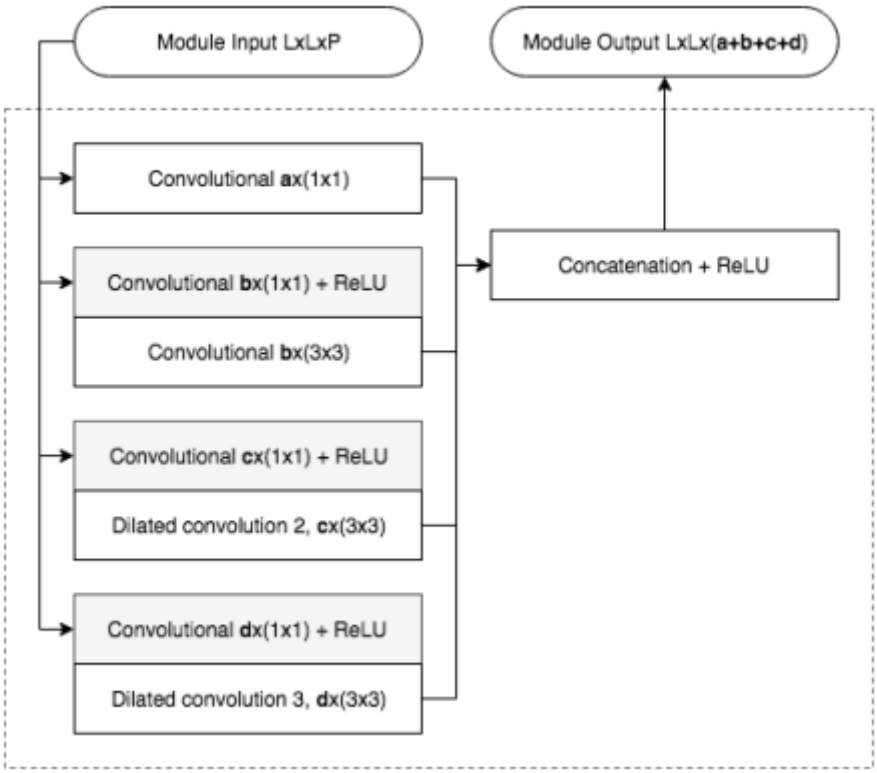

Figure. Architecture of the inception modules used in MesoInception-4. The module is parameterized using a, b, c, d ∈ N. The dilated convolutions are computed without stride.

More than two layers could not be replaced with inception modules, and the categorization did not improve. Table 1 lists the selected parameters (ai, bi, ci, and di) for the module at layer i. This network has a total of 28,615 trainable parameters using those hyper-parameters.

Layer	a	b	c	d
1	1	4	4	1
2	1	4	4	2

Table. Hyper-parameters for the modified Inception modules used in MesoInception-4

Future Work:

By the end of October: Downloading the dataset and testing with different number of layers and to generate hyper-realistic forged Deepfake and Face2Face videos/images.

By the end of November: Evaluate the results on the tested dataset and verify the detection rate on both Deepfake and Face2Face.

IDE used: PyCharm Community Edition 2021.2.1