# MesoNet simulation to detect DeepFake images

Jasia Sanjana
Student, CS, BRAC University.

Abstract
Fake images and videos have become a raging social problem along with technological advancements. The term DeepFake originates from deep learning and fake, meaning, fake images and videos created using deep learning. In this case, we can use deep learning itself to detect them. Such a popular detection network is MesoNet. In this project, MesoNet has been applied to an existing database and the model has detected both the fake and real with good accuracy.

Index terms - DeepFakes, Media Forensics, Face Manipulation, Face Recognition, Databases, MesoNet,

## Introduction:

The concept of deep fake refers to images, audio or video that are fakes that depict events that never occurred but unlike methods of manipulating media in the past like photoshop these deep fakes are created by deep neural networks to be nearly indistinguishable from their real counterparts. Here are two  examples of how realistic and impactful Deep Fake can be in [9] and [10]. The advances in the field of deep fakes are equal parts impressive and alarming. On the upside these technologies with which we can alter media will certainly lead to some classic entertainment, but in the wrong hands this technology can be used to spread misinformation and undermine public trust. This means that as we get better at generating deep fakes we must also get better at identifying them. There are various ways to detect fake facial images but one of the most popular ones is the MesoNet network architecture created by [2]. MesoNet is a convolutional neural network designed to detect defects. The MesoNet is basically a binary classifier since it can give only 2 classes of predictions - real images or fake images. It is adequate for the project.

## Literature review:

R.Tolosana [1] is a brilliant way to get started with DeepFake study. The purpose of the survey article is to compile all the popular techniques of face image manipulation and what researchers have come up with to counter it. The survey is well organized as it starts off by dividing the wide spectrum of facial manipulation category into 4 sectors. This work gives a thorough idea of how these face manipulation techniques are applied, available public datasets, key benchmarks and results of research evaluation. The results were summarized based on each sectors as following -

A. entire face synthesis: this manipulation is done in order to change the entire facial structure from the original and turn into something non-existent which however, achieves a very realistic result. A good reminder from [1] is that these kinds of facial manipulation techniques are originally supposed to be used for good deeds. For example, GAN models can create entire face synthesis and those regenerated images can be used for video game and 3D-modeling industries. Since this model creates an entire fake image, its detection accuracy results can be as high as 100%.
B. identity swap (DeepFakes): This manipulation is the more popular one and it exchanges the face from the original image into another image and creates a fake image. [1] divides it into 2 approaches which are FaceSwap and

DeepFakes. [1] presents different approaches taken to solve this manipulation problem. These are all trained models, which means they do not perform the best when they are thrown in unforeseen circumstances.

C. attribute manipulation: As the name suggests in this manipulation, attributes are added or removed. This form of photo manipulation is very common in this day and age. However, it can be harmful as explained in [6]. [1] mentions the lack of publicly available dataset and standard academic research for this sector of manipulation.

D. expression swap: this form of manipulation replaces an expression of an original photo with another. These too are used as playful photo filter techniques. However, sometimes this can become very harmful, for example, in the case of [7]. [1] has compiled only one database regarding this manipulation technique, despite identity swap evolving much more rapidly.

E. The extensive study done in this survey has been very helpful to understand the overall facial manipulation and other research works done in this sector.

Article [2] is one of the most notable works of DeepFake detection up until now. They use 2 popular forgery techniques: DeepFake and Face2Face. The success rate for their proposed model network, MesoNet, in the previously stated datasets are - 98% and 95% respectively. Another outcome of their finding is that in the case of a human face, the eyes and mouth play a very pivotal role in order to detect fake images. Due to the lack of dataset at that time, [2] has created their own database with about 19,500 images.

[3] works with MesoNet as well but they have used a preprocessing module to keep the performance of MesoNet from dropping. To summarize their method of work, the preprocessing module crops the face from the image and increases discrimination among multi-color channels. Then these preprocessed images go through traditional MesoNet modeling methods. They have used 2 datasets that they have used to verify and validate their model: FcaeForensics++ and Celeb-DF. The accuracy in these 2 datasets are 97.4% and 94.3% respectively.

[5] brings change in order to detect DeepFake. They apply multiple spatial attentional heads so that the model can focus on different parts. They use a block by the name of “textural feature enhancement” which helps find the subtle features. Moreover, they use the attention map to collect the textural features of different levels. Their goal was to find a more optimal way to detect DeepFake than simple binary classification. Another concept they have introduced in their paper is - regional independency loss and attention guided data augmentation strategy. After experimenting with this model architecture on the dataset to verify and validate, they prove how they have achieved state-of-art performance.

Data:

[5] has introduced FaceForensics++ which is a database containing fake facial images. This can be used as a training resource for supervised learning for further research. According to the researchers, this dataset contains manipulated images that have achieved 4 of these state-of-the-art methods: FaceSwap, Face2Face, DeppFakes, NeutralTextures. This database contains more than 500,000 frames containing faces from 1004 videos and contains mostly frontal faces for more accurate detection.

While exploring I found another dataset containing deepfake - [8], celeb-DF. This database contains 590 original videos from YouTube and 5639 corresponding high quality

DeepFake videos. These DeepFakes were generated using an improved synthesis process.

After comparing these two very popular DeepFake datasets, I decided to work with [5] but I chose to work with only the DeepFake part of it since the database is too large. I organized the data in a test/train and validation directory where there were individual folders for DeepFake and Real images. Upon inspection, I found out that the data was well organized, clean and each face had enough frames to train with. The images were mostly clear and had good enough resolution.

Methodology:
I have used the Meso-4 model among the MesoNet model versions. Meso4 is a convolutional neural network with four convolutional blocks followed by one fully connected hidden layer. I made predictions on a portion of the image data from [5], where we examine four sets of images: a) identify deep fakes b) identified reals c) misidentified deep fakes and d) misidentified reals. The python code of meso-4 requires Python 3.5, Numpy 1.14.2, Keras 2.1.5. TensorFlow 3.4.1 and Matplotlib 2.4.1.
After importing all the necessary modules, I created a dictionary called image dimensions to store the image dimensions which are the height and width of the image in pixels and number of color channels. Then I made a classifier. Next a meso4 class is created which takes the classifier class as an argument. Then we proceed to make the network architecture. After that comes the 4 convolutional blocks which includes the convolutional layer and a max pooling layer. with successive blocks in the convolutional base CNNs proceed to higher order feature representations from lines to corners to shapes to faces. MesoNet has four blocks in its convolutional base followed by a fully connected hidden layer and then the output layer for the prediction. With this, the network architecture has been established. Next, it's turn to instantiate the model and load the weights. I downloaded the weights from the MesoNet github repo and then saved the file Meso-4_DF in a folder called weights. And we put the data we have collected and sorted according to our need in the data folder. With this, we are now ready to make a visual presentation of our project.

Results: the result for the dataset looks quite good. The closer the confidence value is to 1, the more likely it is that the image is real. If it was the opposite and the confidence in a real image was below 0.5 then we would have concluded that the model did not predict well.

Further work: when studying about all forms of image forgery a common phrase I noticed was “lack of quality dataset”. Since supervised learning requires a hefty amount of training data for good accuracy and precision, collecting more quality datasets will be beneficial. A fun idea could be collecting data from Bangladeshi old cinema frames to see how the models from [2], [3], [5] work on very noisy images. Another idea can be including few shot and zero shot learning to make the models learn from a very low amount of training data.

Conclusion: DeepFake detection may seem like an endless fight against image forgery but it is a necessary step to take in order to stop it for once and all. With new technologies and improved network architectures, image forgery will become more and more easy. We have to use these advanced technologies and previous studies to build more robust and competitive detection systems. MesoNet is a good detection system thanks to this model being modular and portable. Improvement of MesoNet family models will help detect a diverse amount of data.

References:
[1] R. Tolosana, R. Vera-Rodríguez, J. Fierrez, A. Morales & J. Ortega-Garcia. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. 2020 Inf. Fusion, 64, 131-148.

[2] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630761.

[3] Xia, Z.; Qiao, T.; Xu, M.;Wu, X.; Han, L.; Chen, Y. Deepfake Video Detection Based on MesoNet with Preprocessing Module. Symmetry 2022, 14, 939. https://doi.org/10.3390/sym14050939

[4] A. Rossler, et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019 pp. 1-11. doi: 10.1109/ICCV.2019.00009

[5] H. Zhao, et al., "Multi-attentional Deepfake Detection," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021 pp. 2185-2194.
doi: 10.1109/CVPR46437.2021.00222

[6] “How Photo Editing and Filters Can Harm Your Body Image with Dr. Patrick Byrne” my.clevelandclinic.org. https://my.clevelandclinic.org/podcasts/health-essentials/how-photo-editing-and-filters-can-harm-your-body-image-with-dr-patrick-byrne . (accessed February 17, 2021).

[7] “Facebook lets deepfake Zuckerberg video stay on Instagram” bbc.com. https://www.bbc.com/news/technology-48607673 (accessed 12 June 2019).

[8] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3204-3213, doi: 10.1109/CVPR42600.2020.00327.

[9] BuzzFeedVideo. “You Won’t Believe What Obama Says In This Video!” YouTube. https://www.youtube.com/watch?v=cQ54GDm1eL0 (accessed on Apr 17, 2018).

[10] WashingtonPost. “Mark Zuckerberg ‘deepfake’ will remain online " YouTube. https://www.youtube.com/watch?v=NbedWhzx1rs (accessed on Jun 18, 2019).