Deep Fake Detection using Deep Neural Architectures

Pruthivi Raj Behera
MT20037
pruthivi20037@iiitd.ac.in

Prabal Jain
MT20115
prabal20115@iiitd.ac.in

T G Narayanan
MT20027
narayanan20027@iiitd.ac.in

1. PROBLEM STATEMENT

Deepfake is a technique used to create convincing images/videos by replacing the face of a source person by the face of target person. The results produced using deep fake models are quite advanced. Public figures such as politicians, celebrities, etc are the ones targeted the ones. The deepfake results with such influential figures as subject can prove to be lethal to the society. Moreover, with a vast collection of their images already available online, the deep fake results using them are near indistinguishable. The only ways to distinguish it from the real video is to look for the minute inconsistency in the hair of the person, asymmetry in the face structure, morphed teeth and the background of the face. This bring about the dire need of deep fake detection tools that are well trained with consistent deep learning techniques.

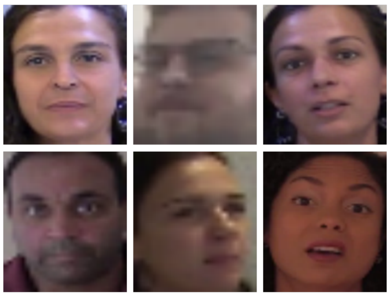


Figure 1: Sample Deepfakes (Top), Sample real (Bottom)

2. LITERATURE SURVEY

Some of the existing works are already done in this area. One of the first works in this field was done by Güera and Delp [2018] with recurrent neural networks. In this paper, the authors created a dataset with 300 videos from video-streaming websites. Additional 300 videos were taken from HOHA dataset to increase the size of dataset. For the method, a Conv-LSTM model was implemented with 80 frames for every video. The channel mean was subtracted from each channel and every frame was resized to 299x299.

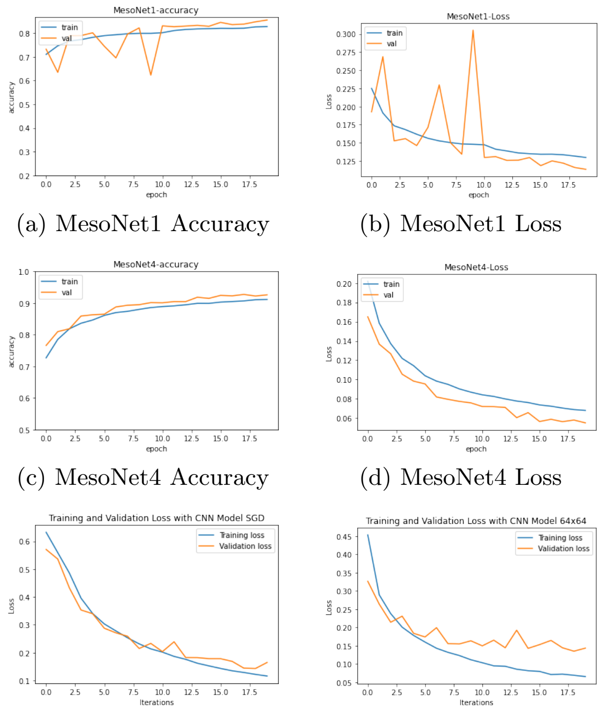
Some more strides were made by Sabir et al. [2019] in the field of recurrent network. In the paper, attempts are made to distinguish the videos prepared using Deepfake, Face2Face or FaceSwap techniques. FaceForensics++ dataset was used for evaluation and a 4.55% increase in accuracy was obtained.

Another recent work was done by Amerini et al. [2019] but with CNN which comprised of optical flow based techniques. In this paper, they have used optical flow fields to distinguish deepfakes from original ones. PWC-Net is a forward flow technique to extract optical flow using the CNN model. A semi-trainable CNN named Flow-CNN is given a input of computed optical flow. They have tested VGG16 and ResNet50 as backbones. The initial layers of network were fixed while last layers including dense one were trained and adam optimizer was used.

Alternative methods such as different ensembling methods have also been tried. One such study was done by Rana and Sung [2020] to address the challenges posed by deepfake multimedia. The methods such as stacking ensemble and randomized weighted ensemble were used where the former uses the output of base-learners to train meta learner for better mapping of results and latter optimizes the weights and take its average.

Most recent work done is done by Guarnera et al. [2020] in this field with convolutional traces. In this paper, a novel deepfake detection method focused on images representing human faces (convolutional traces) was introduced. Different GAN's like STYLEGAN2, FACEFORENSICS++, etc. were used to generate deepfake and classification tests were carried out on the obtained feature vectors with highest accuracies in range of 92-97%.

We will be experimenting the DFDC dataset Dolhansky et al. [2020] since it is by far the most detailed dataset which complies with the ethics of the society. Each person featured in the videos are volunteers or paid actors who are aware of the terms of use of their recordings. The DFDC dataset comprises of 128,154 clips sourced from 3,426 paid actors. Out of this, 104,500 are unique fake videos. Augmentations such as Gaussian blurring, grayscale conversion, horizontal flipping, resolution alteration, rotation, etc were applied to the test set. Various methods such as DFAE, MM/NN Face Swap, NTH, FSGAN and StyleGAN were used. DFAE produced the best results while StyleGAN produced the worst overall results, both at the frame level and at the video level.



(a) MesoNet1 Accuracy

(b) MesoNet1 Loss

(c) MesoNet4 Accuracy

(d) MesoNet4 Loss

(e) CNN Model SGD Loss

(f) CNN Model Adam Loss

Figure 2: Accuracy and Loss Curves

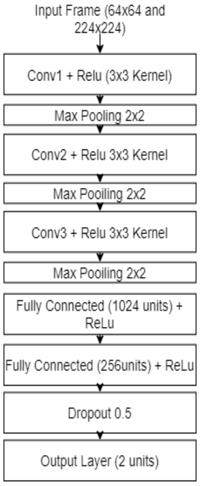


Figure 3: CNN Architecture

Table 1: Classification accuracies and Loss for baseline models performed on DFDC dataset.

Model	Train Loss	Test Loss	Train Acc	Test Acc
CNN-SGD	0.119	0.169	95.09	93.19
CNN-Adam	**0.066**	**0.134**	**97.58**	**95.08**
Meso-1	0.130	0.113	86.03	85.85
Meso-4	**0.070**	**0.055**	**93.38**	**92.52**

3. BASELINES

We have implemented the following 2 variants for each of the 2 baselines models as mentioned below. For implementation, we worked on a subset of the dataset performing undersampling with 20 frames for fake videos and oversampling with 150 frames for real videos to combat class imbalance. We have also tabulated the results as shown in Table 1 and generated plots as shown in Figure 2.

3.1 Deep Convolutional Neural Net

Convolutional neural network (CNN) has been traditionally proven to be effective in fields such as image classification and recognition. The advantage of using CNN is that it can automatically learn complex feature utilizing massive simple neurons and back-propagation. Figure 3 shows the architecture for our deep convolutional neural network in which we used ReLu activation Function with kernel size 3x3, we have used our CNN Model on input frame 64x64 on SGD and Adam optimiser, Also last hidden layer has the keep probability of 0.5 and our output Layer consist of 2 units each units will give the probability of image belonging to that particular class.

3.2 MesoNet

In MesoNet Afchar et al. [2018], the authors presented a method to detect face tampering in videos via deep learning approach. Before processing with Meso-4, the images are resized to the size of 256x256. In Meso-4 architecture as shown in figure 4, the network begins with a sequence of 4 convolution layers followed by a pooling layer. The convolution layers use ReLU activation to improve generalization. Batch normalization layers are used to regularize the outputs and prevent vanishing/exploding gradients. Dropout layers are also used to prevent overfitting and prevent complex computations due to the immense amount of data.

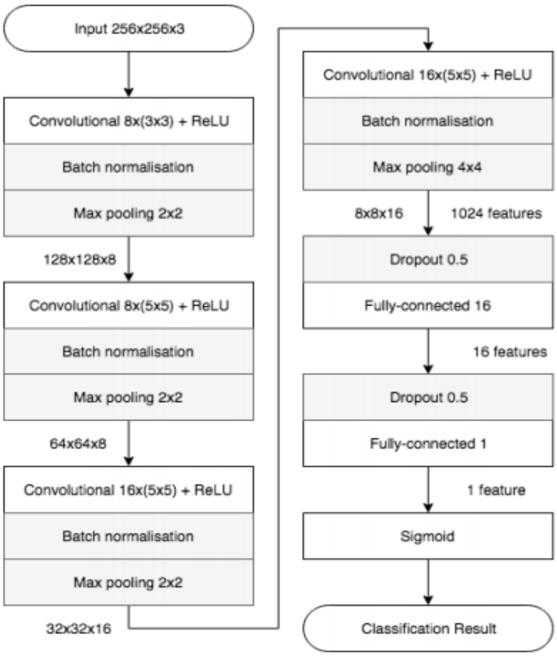


Figure 4: Meso-4 Architecture

References

D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 1(2), 2020.

L. Guarnera, O. Giudice, and S. Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 666–667, 2020.

D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

M. S. Rana and A. H. Sung. Deepfakestack: A deep ensemble-based learning technique for deepfake detection. In *2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, pages 70–75. IEEE, 2020.

E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3(1), 2019.