Image Sequencing for DeepFake Detection

CS3011 Introduction to Artificial Intelligence Project

Team Members:

Harshal Gadhe – 2019068

Kuldeep Singh Gahlot – 2019083

Under Supervision of:

Prof. Kusum Kumar Bharti

1. Introduction

Over the past few years, huge steps forward in the field of automatic video editing techniques have been made. According to several reports, almost two billion pictures are uploaded every day on the internet. This tremendous use of digital images has been followed by a rise of techniques to alter image contents, using editing software like Photoshop for instance. In particular, great interest has been shown towards methods for facial manipulation, face swapping, etc. thus becoming a great public concern recently. On the other hand, these technological advancements open the door to new artistic possibilities (e.g., movie making, visual effects, visual arts, etc.).

In the paper [1] MesoNet: a Compact Facial Video Forgery Detection Network Meso-4 network was proposed but it can become a failure if extracted the wrong frames. This paper addresses the problem of detecting the forged videos created using the DeepFake technique with the help of Image Sequencing.

1.1. Deepfake

Deepfake is a technique that aims to replace the face of a targeted person with the face of someone else in a video. It first appeared in autumn 2017 as a script used to generate face-swapped adult content.

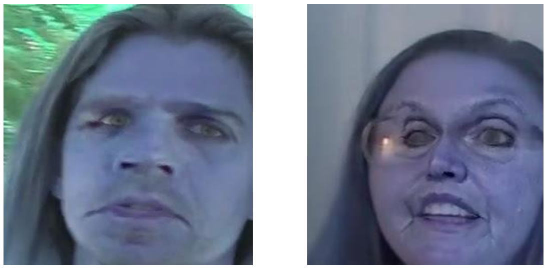


Figure 1. Some of the images extracted from the video were forged using the DeepFake technique. One can easily notice that the images are forged but in the video, there are many such images stack together to create a video and thus it becomes difficult to notice unless one plays the video frame by frame.

1.2. Image Sequencing

An image sequence is a series of sequential still images that represent frames of an animation/video. Using RNN’s these images are then fed to the model altogether and hence the model can predict after considering the whole sequence which can be useful for the prediction of the action of a person, etc.

2. Proposed Method:

This section presents several effective approaches to deal with either DeepFake detection. The proposed method in the research paper [1] MesoNet: a Compact Facial Video Forgery Detection Network was able to achieve an accuracy as high as 90% but since the frames and the gap were defined according to the dataset it may fail in real-time since deepfake can generally be noticed while the person is in motion and thus this approach can be a failure unless we can extract the correct frame and set the optimal gap between the frames which is practically impossible.

We proposed a model by using image sequencing. The image sequence is a method in which the model is fed with a set of images at a time instead of a single image. Feeding the model with a sequence can help the model to use previous images fed and then give a single output to the video. To implement this we build a custom model and used transfer learning (Resnet50 model) along with LSTM layers to increase our accuracy.

During the DeepFake creation, the GAN model focuses on changing the facial

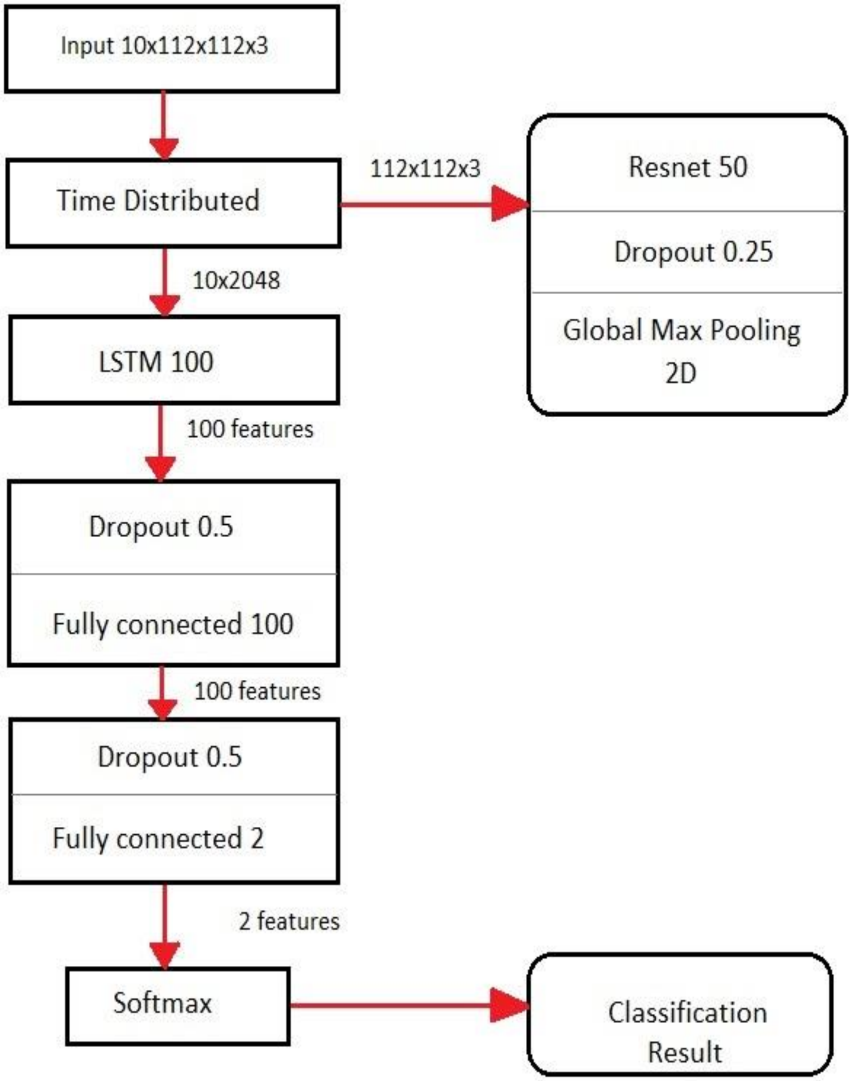


Figure 2. The custom architecture was used to train the model

expressing or swap the faces rather than the body using to our advantage while pre-processing the videos we used face recognition over each frame of a video and cropped the frames having faces in them and by stacking the cropped images we were able to create a video consisting of only faces. After pre-processing the average number of frames in each video was found to be greater than 100. The pre-processed dataset generated was of 5 dimensions and had shapes as (video_files,number_of_frames,112,112,3) [len(dataset) ,no_of_frames, (img_shape), number of channels]. Since the generators used in image pre-processing could handle only a dataset having 4 dimensions we used **VideoFrameGenerator** which can be used for a dataset having 5 dimensions and used the generator to create data having dimensions (video_files, 10, 112, 112, 3).

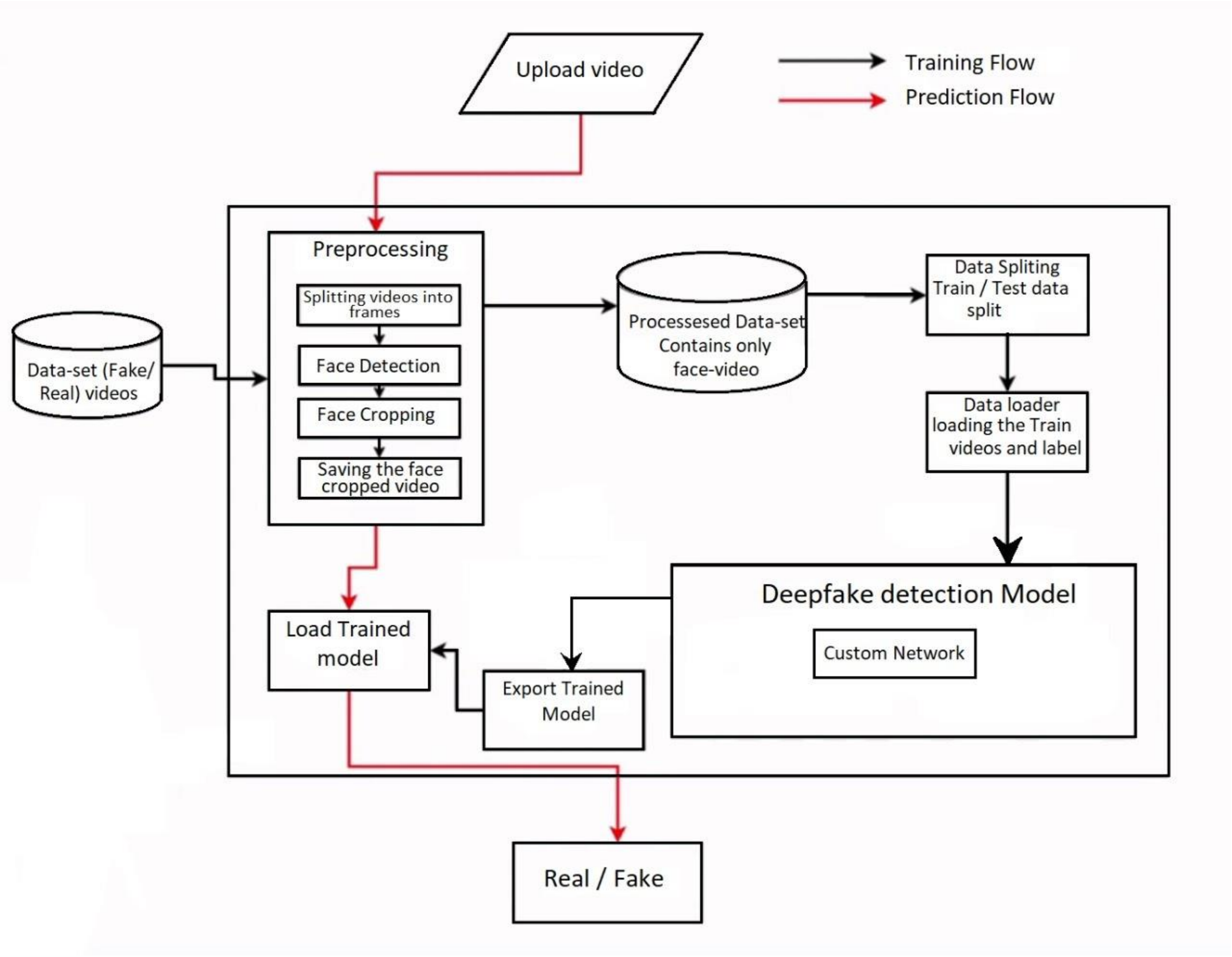

Upload video
Training Flow
Prediction Flow
Preprocessing
Splitting videos into frames
Face Detection
Face Cropping
Saving the face cropped video
Data-set (Fake/ Real) videos
Processesed Data-set Contains only face-video
Data Spliting Train / Test data split
Data loader loading the Train videos and label
Deepfake detection Model
Custom Network
Export Trained Model
Load Trained model
Real / Fake

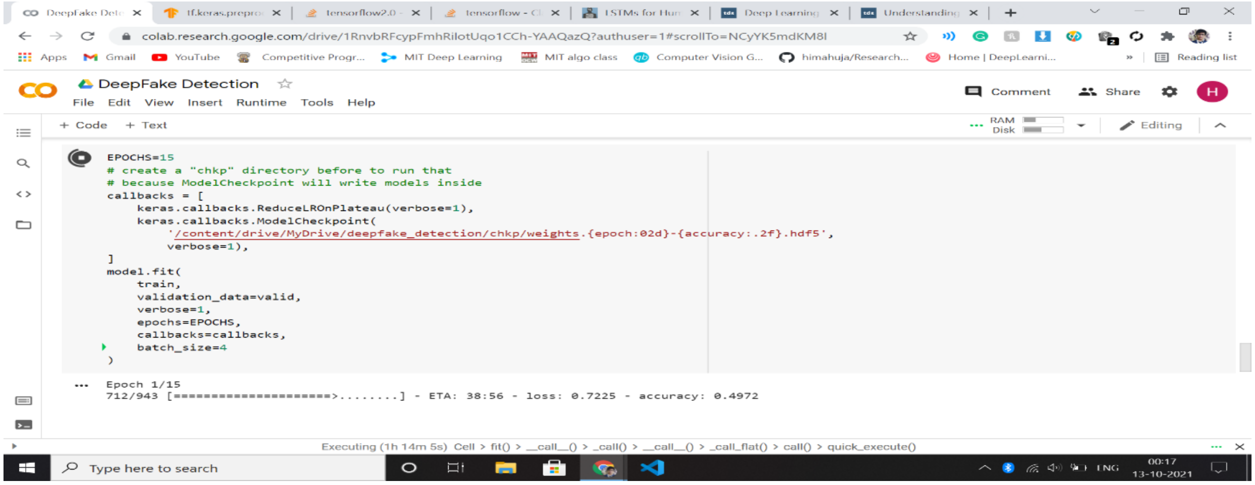
Figure 3. Training and Prediction Pipeline flow

The Dataset used in the proposed model consisted of a mixed dataset of face_forensics, celeb_dataset, deepfake detection challenge, and hence it can work better in real-time than the model used in [1] which was trained only on either DeepFake or Face2Face dataset.

During the prediction of unseen data, we will pass the video to the pre-processing function where we have predefined the frames which we will extract from the video for the prediction. The frames to be selected are taken in order as 5,10,15….100 from the 20 frames extracted we will only keep 11 of them which have face in them and using the MTCNN face recognition package we will crop the area containing the faces. Later these 11 frames will be pass to the model for prediction and the results of each frame are recorded in the array and then the average is taken for each class. The label is then predicted as the class having the maximum count.

3. Results

Due to a lack of computational resources / GPU, we were unable to pass all the 100 frames as a sequence to our model and had to pass only 10 frames with only being able to train for 2 epochs after which we exhausted our resources. Resulting in very poor accuracy of our model of over just 50% on both training and validation datasets.



```
EPOCHS=15
# create a "chkp" directory before to run that
# because ModelCheckpoint will write models inside
callbacks = [
    keras.callbacks.ReduceLROnPlateau(verbose=1),
    keras.callbacks.ModelCheckpoint(
        '/content/drive/MyDrive/deepfake_detection/chkp/weights.{epoch:02d}-{accuracy:.2f}.hdf5',
        verbose=1),
]
model.fit(
    train,
    validation_data=valid,
    verbose=1,
    epochs=EPOCHS,
    callbacks=callbacks,
    batch_size=4
)
```

```
Epoch 1/15
712/943 [=====================>........] - ETA: 38:56 - loss: 0.7225 - accuracy: 0.4972
```

```
Executing (1h 14m 5s) Cell > fit() > __call__() > _call() > __call__() > _call_flat() > call() > quick_execute()
```

Figure 4. Model while training

4. Conclusion

The model generally considers the blur area or abnormal facial expression around the face to predict the image as fake but it can also be caused due to motion of the person which can result in the wrong prediction. Image Sequencing as in the proposed model takes multiple images at a time and used RNN's for its prediction thereby considering the motion of the motion hence classifying the video more accurately than its counterparts. Similar pre-processing was used

5. References :

1. Deep Learning for Deepfakes Creation and Detection: A Survey Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Cuong M. Nguyen, Dung Nguyen, Duc Thanh Nguyen, Saeid Nahavandi, Fellow, IEEE(April 2021)

2. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection, International Research Journal of Engineering and Technology (2020)

3. MesoNet: a Compact Facial Video Forgery Detection Network (2018)

4. Video Face Manipulation Detection Through Ensemble of CNNs, International Research Journal of Engineering and Technology (April 2020)

5. Unmasking DeepFakes with simple Features (March 2020)

6. Reverse Engineering of Generative Models: Inferring Model Hyperparameters from Generated Images (June 2021)