MAR ATHANASIUS COLLEGE OF ENGINEERING

KOTHAMANGALAM

Department of Computer Applications

Initial Project Report

Video Face Manipulation Detection Through Ensemble of CNNs

Done by

Nandu Sasikumar

Reg.No: MAC21MCA-2023

Under the guidance of

Prof. Biju Skaria

2021-2023

1. Abstract

In the last few years, several techniques for facial manipulation in videos have been successfully developed and made available to the masses (i.e., FaceSwap, deepfake, etc.). These methods enable anyone to easily edit faces in video sequences with incredibly realistic results and a very little effort. Despite the usefulness of these tools in many fields, if used maliciously, they can have a significantly bad impact on society (e.g., fake news spreading, cyber bullying through fake revenge porn). Even though the technology was developed by experts, now it is available as web application and mobile application which enables normal people to access it and make manipulated videos and photos easily. Then it has become a big threat to the society. The ability of objectively detecting whether a face has been manipulated in a video sequence is then a task of utmost importance. One of the points to select this topic as the project is this relevance of this particular problem in the society.

This project is intended to detect deepfake manipulation in videos. And the entire project is completely based on the paper titled “Video Face Manipulation Detection Through Ensemble of CNNs”, which was presented at ICPR2020. The proposed method is based on the concept of ensembling. Indeed, it is well known that model ensembling may lead to better prediction performance. I therefore focus on investigating whether and how it is possible to train different CNN-based classifiers to capture different high-level semantic information that complement one another, thus positively contributing to the ensemble for this specific problem. The proposed method in the paper takes inspiration from the family of EfficientNet models and improves upon a recently proposed solution, investigating an ensemble of models trained using two main concepts: (i) an attention mechanism (ii) a triplet siamese training strategy. The dataset that I wish to use is one that provided in the site Kaggle as part of a competition.

[1] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini and S. Tubaro, "Video Face Manipulation Detection Through Ensemble of CNNs," 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 5012-5019, doi: 10.1109/ICPR48806.2021.9412711. - IEEExplore

[2] Karandikar, Aarti. (2020). Deepfake Video Detection Using Convolutional Neural Network. International Journal of Advanced Trends in Computer Science and Engineering. 9. 1311-1315. 10.30534/ijatcse/2020/62922020. - ResearchGate

[3] M. S. Rana and A. H. Sung, "DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection," 2020 7th IEEE International Conference on Cyber Security and loud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), 2020, pp. 70-75, doi: 10.1109/CSCloud-EdgeCom49738.2020.00021. - IEEExplore

2. Literature Review

[1] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini and S. Tubaro, "Video Face Manipulation Detection Through Ensemble of CNNs," 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 5012-5019, doi: 10.1109/ICPR48806.2021.9412711. - IEEExplore

This paper is the base paper for my project. In this paper, the authors tackle the problem of face manipulation detection in video sequences targeting modern facial manipulation techniques. In particular, they study the ensembling of different trained Convolutional Neural Network (CNN) models. In the proposed solution, different models are obtained starting from a base network (i.e., EfficientNetB4) making use of two different concepts: (i) attention layers; (ii) siamese training. Combining these networks leads to promising face manipulation detection results on two publicly available datasets with more than 119000 videos. The proposed method takes inspiration from the family of EfficientNet models and improves upon a recently proposed solution, investigating an ensemble of models trained using two main concepts: (i) an attention mechanism which generates a human comprehensible inference of the model, increasing the learning capability of the network at the same time; (ii) a triplet siamese training strategy which extracts deep features from data to achieve better classification performances.

[2] Karandikar, Aarti. (2020). Deepfake Video Detection Using Convolutional Neural Network. International Journal of Advanced Trends in Computer Science and Engineering. 9. 1311-1315. 10.30534/ijatcse/2020/62922020. - ResearchGate

This paper aims to solve the problem by proposing a model that analyses the frames of the videos using deep learning approach to detect inconsistencies in facial features, compression rate and discrepancies introduced in the videos while creating them. The model uses a convolutional neural network along with transfer learning to train the model that can catch these instilled errors in the deepfakes. The neural network is trained on these discrepancies induced during deepfake creation around the face. It uses a dataset called “Celeb-DF: A New Dataset for DeepFake Forensics” to train the model. The paper further discusses methods that can be used, in detail, to improve learning by this model. In this paper, proposed method uses transfer learning on VGG-16 model to train the dataset and focus on facial manipulation for detection of forgery. Transfer learning is essential as the model should be trained in considerable amount of time and should require minimum resources to give the desired accuracy for its classification over varied examples in the dataset. The proposed model works well and is able to successfully

gather features required for further processing to test for deepfakes. For improving the performance, further research can be done on detecting temporal and audio discrepancies and then using this combined information with features extracted from image processing module. It is observed that the accuracy of the proposed model decreases with low quality images and with medium quality videos the accuracy needs to be further increased using combined models for training on temporal parameters. Thus, better dataset with improved quality will lead to better training. Various ensemble learning techniques can also be implemented to further increase the accuracy of the model and account for variance in the dataset. Aggregation of results over each frame and over different learning models will thus give best results. The authors hope that the presented techniques for analysis on deepfakes will pave the way for further research in the field of image & video forgery and digital media forensics.

[3] M. S. Rana and A. H. Sung, "DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection," 2020 7th IEEE International Conference on Cyber Security and loud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), 2020, pp. 70-75, doi: 10.1109/CSCloud-EdgeCom49738.2020.00021. - IEEExplore

In this paper, the authors propose a deep ensemble learning technique, DeepfakeStack, by experimenting with various DL-based models on the FF++ dataset. The experiment shows that a larger stacking ensemble neural network (called DFC) model is defined and fit on the test (unseen) dataset, then the new model is used to predict the test dataset. Evaluating the results, they see that the proposed DFC model achieves a good accuracy and AUROC 1.0, outperforming the DL-based models, thereby provides a strong basis for developing an effective Deepfake detector.

3. Problem Definition

Deep Learning Algorithm development has advantages and downsides of its own. New advanced AI techniques are being utilised to produce false videos as a result of new technological advancements in AI. Such videos might be utilised maliciously and constitute a serious danger to society in a variety of social and political contexts. Deepfakes are the name for these fake videos. Deepfakes are artificial intelligence produced video manipulations or other digital representations that generate faked sounds and pictures that seem real. A deep-learning system may create a convincing copy by examining several photos and videos of a target individual, then imitating their behaviour and voice patterns. Because more realistic deepfake production technologies

are always being developed, it is extremely difficult to detect these videos. The proposed model will be able to detect deepfake manipulation in videos.

4. Exploring the Dataset

➢ Source

Deepfake Detection Challenge
https://www.kaggle.com/c/deepfake-detection-challenge/data

➢ Description

The dataset that I wish to use is one that provided in the site Kaggle as part of a competition - Deepfake Detection Challenge. The full dataset is about 400 GB but there is also a smaller dataset available with 800 videos that of 4.4 GB. The data is comprised of .mp4 files. A metadata.json accompanies each set of .mp4 files, and contains filename, label (REAL/FAKE), original and split columns, listed below under Columns.

➢ Feature List and Class

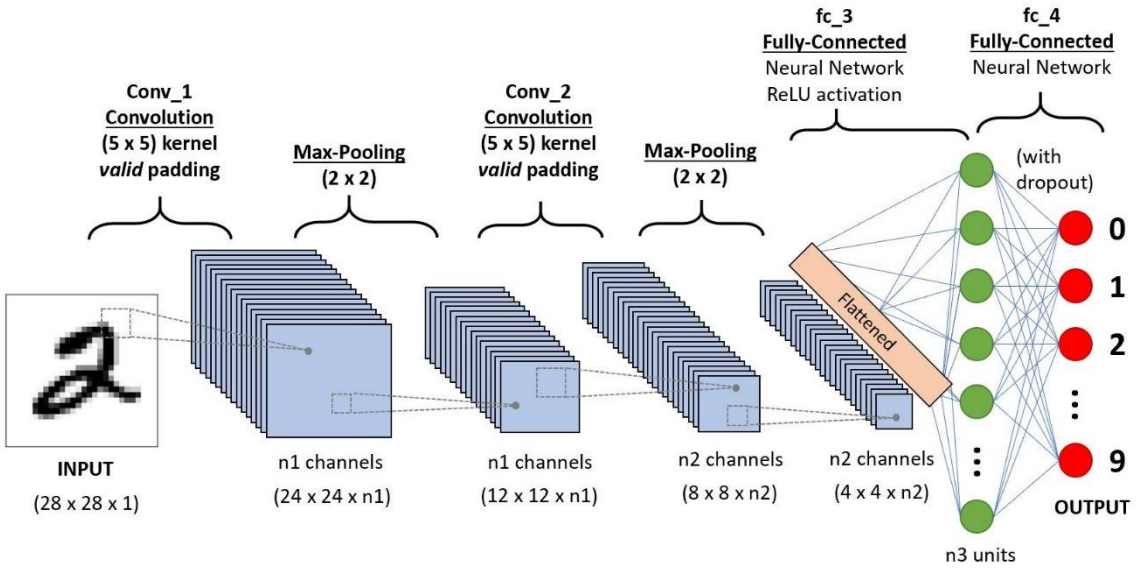
The Facebook DeepFake Detection Challenge (DFDC) Dataset is part of the DeepFake detection challenge, which has 4,113 DeepFake videos created based on 1,131 original videos of 66 consented individuals of various genders, ages and ethnic groups. The data is comprised of .mp4 files. All videos are of length under 10 seconds. The dimension of video files are 1920 X 1080 and an approximate frame rate of 30.

train_sample_videos - a folder containing a sample set of training videos and a metadata.json with labels.
sample_submission.csv - a sample submission file in the correct format.
test_videos - a folder containing a small set of videos to be used as a public validation set.
columns

- filename - the filename of the video
- label - whether the video is REAL or FAKE
- original - in the case that a train set video is FAKE, the original video is listed here
- split - this is always equal to "train"

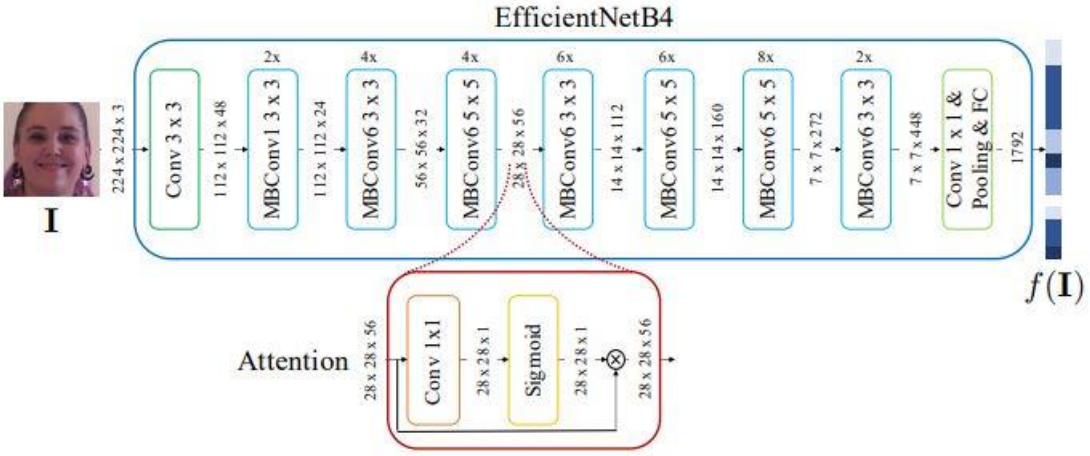
5. Architecture

The proposed method is based on the concept of ensembling. Indeed, it is well-known that model ensembling may lead to better prediction performance. Therefore, focuses on investigating whether and how it is possible to train different CNN-based classifiers to capture different high-level semantic information that complement one another, thus positively contributing to the ensemble for this specific problem.



General CNN architecture

I consider, as starting point the EfficientNet family of models. Among the family of EfficientNet models, I choose the **EfficientNetB4** as the baseline for my work, motivated by the good trade-off offered by this architecture in terms of dimensions (i.e., number of parameters), run time (i.e., FLOPS cost) and classification performance. Then explicitly implement an attention mechanism similar to the one already exploited by the EfficientNet itself, as well as to the self-attention mechanisms. On one hand, this simple mechanism enables the network to focus only on the most relevant portions of the feature maps, on the other hand it provides us with a deeper insight on which parts of the input the network assumes as the most informative. Indeed, the obtained attention map can be easily mapped to the input sample, highlighting which elements of it have been given more importance by the network. The result of the attention block is finally processed by the remaining layers of EfficientNetB4. The whole training procedure can be executed end-to-end, and we call the resulting network **EfficientNetB4Att.**



Attention mechanism

Applying siamese training on these models we get the models **EfficientNetB4ST** and **EfficentNetB4AttST** respectively. I have also used the baseline network **XceptionNet**.

6. Project Pipeline