Deep Fake Detection Through Convolutional Neural Network

Gina Hamdy[1] , Esraa Ali[1] , Manar Saad El-Din[1] , Yousef Adel[1] , Ahmed Mohamed Ahmed Hussien[1] , Ahmed Yousef El Sayed[1] , Mahmoud Ali[1] ,Ghada Nady[1], Yasser Omar[2]

[1]Modern University for Technology and Information
[2]Arab Academy For Science , Technology And Maritime Transport

Abstract

Fake media have been defined as the videos/images, but without running any types of tests to validate the truth of this content. Fake Media and faux accounts represent a really critical issue whose complexity grows in a common place. In recent years, topics such as fake media and accounts detection have received a lot of attention within the research fields. Therefore, using and applying a variety of Deep learning methods and neural networks to effectively detect fake media is the main goal of the paper . Thus, CNN networks have been used to perform deepfake detection, with the best results obtained. In this study, The model used is a sequential convolutional neural network combined with other methods as adam optimizer and max pooling with 3 different gathered datasets (Celeb-DF and Faceforensics++). The model ended up with accuracy 93.3% and loss rate 19.5%.

1.Introduction

Over the past few years, huge steps forward in the field of automatic video editing techniques have been made. In particular, great interest has been shown towards methods for facial manipulation.

Just to name an example, it is nowadays possible to perform facial reenactment, i.e., transferring the facial expressions from one video to another one . This enables one to change the identity of a speaker with very little effort. Systems and tools for facial manipulations are now so advanced that even users without any previous experience in photo retouching and digital arts can use them. Indeed, code and libraries that work in an almost automatic fashion are more and more often made available to the public for free [1]. The increased complexity of mobile camera technology , As well as the ever-expanding reach of social media and media sharing sites, have made creating and disseminating digital movies easier than ever before [2]. The absence of advanced editing tools, the high demand for topic expertise, as well as the difficult and time-consuming procedure required have all limited the quantity of false videos and their degrees of realism until recently. However, because of the availability of large-volume of the training data and high-throughput processing, the time spent fabricating and manipulating videos has decreased dramatically in recent years. These technologies pose significant threats to the reliability of visual information, and can represent harmful tools to undermine the digital identity and reputation of

individuals. The many cases of abuse reported in the last months involving public figures in politics and economics, confirm these concerns, and can only expect this phenomenon to increase in the upcoming years. As a response, the detection of the employment of new efficient techniques for synthetic media generation has drawn many research efforts in the last years. An ever increasing number of tools and approaches have been proposed in the last years, together with the development of benchmark datasets (e.g., FaceForensics++) and world-wide open challenges (e.g., Facebook Deepfake Detection Challenge) [3]. Nowadays, it is becoming increasingly easy to automatically synthesize non-existent faces or manipulate a real face (a.k.a. bonafide presentation) of one subject in an image/video, thanks to: *(i)* the accessibility to large-scale public data and *(ii)* the evolution of deep learning techniques that eliminate many manual editing steps such as Autoencoders (AE) and Generative Adversarial Networks (GAN). As a result, open software and mobile applications such as ZAO and FaceApp have been released, opening the door to anyone to create fake images and videos, without any experience in the field [4]. In this paper sectionsI disccussed about background of deep learning methods. SectionII is about literature review on related work. SectionIII explained methodology of how the model works. SectionIV is the result that came from the model. SectionV is the conclusion.

2.Background

Deep learning is a machine learning technique based on the neural network concept. The term "deep" in deep learning refers to the usage of numerous hidden layers in the network. The deep learning architecture, which was inspired by artificial networks, uses an unbounded number of hidden layers of bounded size to extract higher information from raw input data. The number of hidden layers is governed by the training data's complexity. To efficiently deliver the correct results, more complicated data need more hidden layers. Deep learning has been successfully applied in a variety of fields in recent years, including computer vision, audio processing, automatic translation, and natural language processing. When compared to machine learning methodologies, deep learning gives state-of-the-art results in several fields..Deep learning has also shown promise in detecting deepfakes. Several deep learning algorithms have been proposed in the literature, including: 1) convolutional neural network (CNN); 2) recurrent neural network (RNN). These strategies are briefly described in the following sections, followed by an explanation of how they are applied to deepfake discovery.

2.1.Convolutional Neural Network (CNN)

The most common deep neural network model is a convolutional neural network (CNN). CNNs have an input and output layer, as well as one or more hidden layers, similar to neural networks. The inputs from the first layer are read by the hidden layers,which then apply a convolution mathematical process on the input values. Convolution denotes a matrix multiplication or other dot product in this context. Following matrix multiplication, CNN employs a nonlinearity activation function like the Rectified Linear Unit (RELU), followed by additional

convolutions such as pooling layers. Pooling layers' main purpose is to minimize the dimensionality of data by computing outputs using functions like maximum pooling or average pooling [5].

2.2.Recurrent Neural Network (RNN)

Another application of artificial neural networks is the recurrent neural network (RNN), which can learn characteristics from sequence data. RNN is made up of numerous invisible layers, each with a weight and bias, similar to neural networks. Relationships between nodes in a direct cycle graph that run in order in RNN. RRN has the advantage of allowing temporal dynamic behavior to be discovered [6]. RNNs, unlike feed forward networks (FFNs), use an internal memory to remember sequences of information from earlier inputs, making them helpful in a range of applications, such as natural language processing and audio recognition. A temporal sequence can be handled by an RNN by introducing a recurrent hidden state that captures interdependence across time scales.

2.3.Generative Adversarial Network (GAN)

A deep learning model that combines generative and discriminative models,was proposed in 2014 .The generative model may generate data at random ,whereas the discriminative model determines whether or not the gendered data is from training datasets .The competition between the generative and discriminative models may improve GAN in achieving better results . GAN is frequently used in image categorization . picture and text generation , and other tasks

3.Literature Review

At this point an overview about fakeMedia detection papers that discussed most techniques for detecting fake media and explained their perspective models used as well as similar application and related work will be presented

Ekraam Sabir and etel [7],In the model and face alignment approach improves upon the state-of-the-art. Video-based face manipulation became available for the community with the recent release of FaceForensics , followed by its extended and improved version: FaceForensics++. FaceForensics released Face2Face manipulation FaceForensics++ (FF++) is an extension of FF, further augmenting the collection with Deepfake and FaceSwap manipulations. The set comprises 1,000 videos organized into a single split where 720 videos are reserved for training and 140 videos used for validation.The overall approach for manipulation detection is a two step process: cropping and alignment of faces from video frames, followed by manipulation detection over the preprocessed facial region using
CNN + RNN. Ended up with 93% for the accuracy of the model.

Shruti Agarwal and etel [8]. TheA2V synthesis technique takes as input a video of a person speaking and a new audio recording, and synthesizes a new video in which the person's mouth is synchronized with the new audio.

Explored if a more modern learning-based approach can outperform the hand-crafted profile feature. Specifically, trained a convolutional neural network (CNN) to classify if a mouth is open or closed in a single video frame. The input to the network is a color image cropped around

the mouth and rescaled to a 128×128 pixels (Figure 1). The output, c, of the network is a real-valued number in [0,1] corresponding to an “open” (0) or “closed” (1) mouth. Ended up with 96.4% for the accuracy of the model.

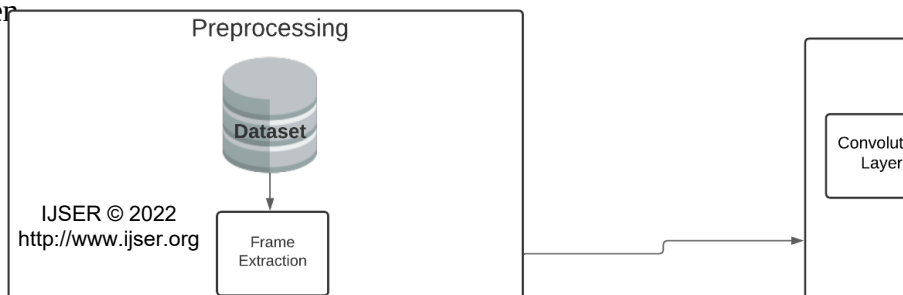
Davide Coccomini and etel..[9] , They conducted the tests on FaceForensics++ and on the 5000 test videos made available for the DFDC dataset. In order to compare their methods also on the DFDC test set, tested the Convolutional Vision Transformer of Wodajo and Atnafu [2021] on these videos obtaining the necessary AUC and F1-score values for comparison.Trained the networks on 220,444 faces extracted from the videos of DFDC training set and FaceForensics++ training videos, and used 8070 faces for validation from DFDC dataset.Convolutional Cross ViT uses two distinct branches: the S-branch, which works on small patches and therefore has a local view of the input image, and the L-branch, which works on larger patches and therefore has a more global view. In each of these branches, the Transformer Encoders outputs are combined through cross attention, which allows a direct interaction between the two results. Ended up with 87% for the accuracy of the model.

4.Proposed Model

The first section of the model is preprocessing where it focuses mainly on the face by firstly framing video , then face detection , face cropping and alignment. Then it takes a path to the processed dataset which is celeb (v1 , v2) to begin the training phase on the samples inside it. Lastly , the manipulation detection phase by extracting features using CNN. Then loading a trained model and deciding whether it's fake or real. As it is illustrated in fig 1.

4.1.Datasets

The DeepFake Forensics (Celeb-DF) dataset, which contains synthetic videos of higher visual quality, to better test existing DeepFake detection algorithms and to support the development of more effective detection systems.DeepFake films in these datasets feature a variety of visual artifacts that make them easily distinguishable from actual videos.The Celeb-DF dataset contains 590 genuine films and 5,639 DeepFake movies (for a total of almost two million video frames). The average length of all videos is about 13 seconds, with a frame rate of 30 frames per second. The genuine videos are drawn from publicly available YouTube recordings and correlate to interviews with 59 celebrities of various genders, ages, and ethnicities. In the true videos, 56.8 percent of the individuals are male, while 43.2 percent are female. 8.5 percent are over the age of 60, 30.5 percent are between the ages of 50 and 60, 26.6 percent are in their 40s, 28.0 percent are in their 30s, and 6.4 percent are under the age of 30.There are 5.1 percent Asians, 6.8 percent African Americans, and 88.1 percent Caucasians. The second dataset is FaceForensics++ is a forensics dataset consisting of original video sequences that have been manipulated with four automated face



manipulation methods: Deep Fakes, Face2Face, FaceSwap and NeuralTextures. As providing binary masks the data can be used for image and video classification as well as segmentation. In addition, it provide 1000 Deep Fakes models to generate and augment new data.

Deepfake		
Dataset	Real	Fake
Celeb-DF v1	408	795
Celeb-DF v2	890	5,639
FaceForensics	1000	1000

Table 1. Dataset

4.2.Preprocessing

Used computer vision library to make resize to crop the frame , detect the face and alignment to the center of the frame. Then , Normalize the frame to see which data from each pixel to be taken and which to be neglected.

4.3.Feature Extraction

Firstly ,the dataset is splitted to two sections (Real or fake) then resized to all frames to avoid unnecessary computations. The maximum number of training images will be 8000. if the number of training images increases above 8000 it will get higher results. But this will be computationally expensive.

4.4.CNN Model

Model is sequential and divided into two convolution layers. The two layers use 64 filters , kernel size (3,3) and activation function is The rectified linear activation (ReLU). First layer takes the image shape from the feature extraction. Then, batch normalization to stabilize the learning processing and to reduce the training epochs. Max pooling is used to split the frames by matrix (2,2) the global average pooling that calculates the average outputs of even feature.And the final method is

dense for changing the dimensions of the vectors by using every 265 neuron by ReLU after that used softmax for the output layer of neural network.

5.Experiment Result

CNN		
Manipulation	Accuracy	Loss
Deepfake (with 50 epoch)	0.9334	0.1958

Table 2. Accuracy and Error rates

Additionally, Report the areas under the receiver operating curve (AUC) scores. All numbers are reported on Celeb-DF(v1,v2). Using for training Adam optimizer with 1e-4 learning rate. Additionally, all results are on the heavily compressed version of the dataset. It does not evaluate high and low quality videos since the baseline performance for those is already very high. Table 2 shows our results from the choice of face alignment method and multiple levels of recurrence of our model. For the simpler.specifically , a simple CNN with 2 convolution and max-pooling layers each followed by a feedforward network as the localization net and bilinear interpolation for the sampler.Specifically, since the DenseNet used in our experiments has two blocks for generating feature maps.

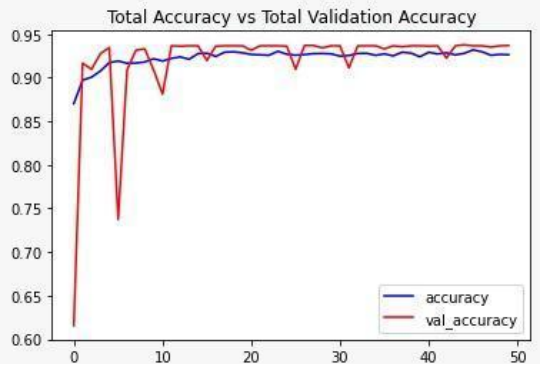


Fig 2 Accuracy curve of CNN performance

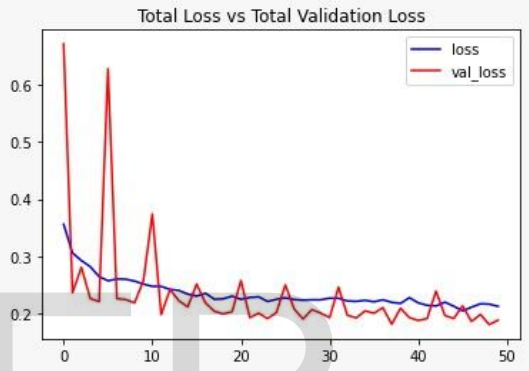


Figure 3.Loss curve of CNN performance

6.Conclusion

This study demonstrated the effectiveness of deep neural network techniques in deepfake detection criteria. Presented an approach that can automatically detect deepfake based on deep learning concepts. Information sharing among feature extraction, preprocessing and model building tasks improved the model overall performance.The model provided good level accuracy and reliability. In the near future one can extend this work by exploring more architectures that will help in implementing new detection techniques to detect deepfakes.

7.References

[1] Bonettini, N., & Cannas, E. D. Video Face Manipulation Detection Through Ensemble of CNNs. (2020).

[2] Aditi Kohli & Abhinav Gobta. Detecting Deepfake, Faceswap and Face2Face facial forgeries using frequency CNN.(2020)

[3] Marcon, F., Pasquini, C., Boato, G.& etel. Detection of Manipulated Face Videos over Social Networks: A Large-Scale Study.(2021)

[4] Tolosana, R., Rodriguez, R. V.-.. An Introduction to Digital Face Manipulation.(2022)

[5] Pishori, A., Rollins, B., & etel... Detecting Deepfake Videos: An Analysis of Three Techniques. (2020)

[6] Heo, Y-J, Kim, B.-G, & etel… Deep Face Detection scheme based on vision transformer and distillation.(2021)

[7] Sabir, E., Cheng, J., "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos.".(2021)

[8] Agarwal, S., & Fried, O. & etel.. .Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. (2021)

[9] Coccomini, D., Messina, N. & etel.. COMBINING EFFICIENTNET AND VISION TRANSFORMERS FOR VIDEO DEEP FAKE DETECTION. (2021)

.