



AI FAKE AND REAL FACE DETECTION

A Phase I Report on Developing a scalable Detective model for Fake or Real Face Detection using Deep learning

¹Greeshma P, ²Shashidhar Kini k,

¹Student, ²Professor and Head,

¹Department of Master of Computer Applications,

¹Srinivas Institute of Technology valachil, Mangalore, Karnataka, India

Abstract: Fake images are one of the most widespread phenomena that have a significant influence on our social life, particularly in the world of politics and celeb. Nowadays, generating fake images has become very easy due to the powerful yet simple applications in mobile devices that navigate in the social media world and with the emergence of the Generative Adversarial Network (GAN) that produces images which are indistinguishable to the human eye. This makes fake images and fake videos easy to perform, difficult to detect, and fast to spread. As a result, image processing and artificial intelligence play an important role in solving such issues. Thus, detecting fake images is a critical problem that must be controlled and to prevent these numerous harmful effects. This research proposed utilizing the most popular algorithm in deep learning is (Convolution Neural Network) to detect the fake images. The first steps includes a preprocessing which start with converting images from RGB to Cyborg color space, after that entering the Gamma correction. finally extract edge detection by entering the Canny filter on them. After that, utilizing two different method of detection by applying (Convolution Neural Network with Principal Component Analysis) and (Convolution Neural Network without Principal Component Analysis) as a classifiers. The results reveal that the use of CNN with PCA in this research results in acceptable accuracy. In contrast, using CNN only gave the highest level of accuracy in detecting manipulated images.

Index Terms- Deep learning, Deep fake, Generative Adversarial Network, Convolution Neural Network, Principal Component Analysis.

I. INTRODUCTION

Due to the technological development and the wide spread of programs for deep fakes on digital images with the advent of Generative Adversarial Network (GAN). Image and video editing are becoming easier and considered a cybercrime due to the fake information and images can be spread fast and widely in Internet through social media. Thus, deepfake can be considered as the ability to automatically create, modify, or swap a person's face in videos and images using algorithms depending on Deep Learning technology, that is one of the phenomena that is expanding quickly. It is feasible to produce top-notch results by developing new multimedia materials that are difficult for the human eye to distinguish between real and fake. This term "Deepfake" refers to all multimedia items that have been synthetically modified or produced using generative machine learning models [1]. DeepFake is concerning because it combines a high level of authenticity, quick evolution and pervasiveness [2]. On Reddit, DeepFake has been widely utilized as of November 2017. due to the United States widespread manufacturing of pornographic videos that has gained a solid online reputation and attracted interest of the people from all walks of life.

In January Deep Fake was officially used in an application in 2018. As a result DeepFake proliferation was accelerated. The object of face swapping has also grown in popularity from celebrities and politicians to students, friends and coworkers As a result, people's panic were naturally prompted. The main IT companies have also begun to collaborate action with the academic community to avoid additional detrimental impact on the fierce discussion about whether and how to control Deepfake technology [3]. With the continuous enhancement of the computing power of computer, the ongoing reduction of hardware price and the high integration of deep learning tools like tensor flow [4] and keras [5], technicians with specific professional backgrounds can produce high-quality faked images and videos consistent with the real distribution of data through the use of convolution automatic encoder [6] and Generative adversarial networks (GAN) [7].

II. LITERATURE SURVEY

1. Y. Wang et al in 2021 presented two algorithms for detection of fakeface images. The first approach is the Local Binary Pattern (LBP)-Net using global texture features used to detect fakefaces. The second method ensemble model constructed from five models including LBP-Net, Gram-Net, ResNet and two models utilizing InceptionResnetV1 pre-trained on Casia-Webfaceare and vggface2. Results of detecting fakeface images by several image augmentation such as downsample (66.32%), brightness

(81.09%), Solarize (75.04%), Contrast (85.42%) and color (91.06%) when using “140K Real and Fake Faces” [9].

2. M. Taeb et al in 2022 compared the most popular state-of-the-art face-detection classifiers such as CNN, VGG19, and DenseNet-121 using an enhanced actual and fake face dataset. Data augmentation is a technique for improving performance with conserving computing resources. When compared to other studied models, early findings show that VGG19 has the best performance and accuracy of 95%. When using “140K Real and Fake Faces” [10].

3. S. Tariq and et al. in 2018 proposed an ensemble of three different convolutional neural network with different layers as a classifier after performing pre-processing represented by face cropping and noise filtering methods. Finally a fully automated end-to-end fakeface detection pipeline developed to be focused on image content with only RGB channel information in order to recognize GAN-created face images with 94% AUROC score (Area Under the Receiver Operating Characteristic curve) and recognize human-created fake face images with 74.9% AUROC score. Where, applied on CelebA and PGGAN dataset images. [11].

4. X. Chang et al in 2020 improved VGG network termed NA-VGG, it is used to detect deepfake face images, which was based on image noise and augmentation of image. The SRM filter layer is utilized to highlight the noise features of the image and after that the image noise is acquired as the network's input. Second, the image noise map is augmented to make the face features appear weaker. Finally, the augmented noise of images are fed into the network, which is trained and used to determine whether or not the image is fake. The obtained accuracy is 85.7% when using the Celeb-DF dataset [12].

III. DATA SET

Deepfake detection systems often use binary classifiers to group information into fake and real classes. This strategy needs a great quantity of high-quality authentic and manipulated data to train the models of classification. The dataset was taken from kaggle website. This dataset contains of all 70k real faces from of the Flickr dataset gathered by Nvidia, in addition 70k fake faces picked from the Bojan's 1 Million fake faces (produced using StyleGAN). In this dataset, combined both dataset, scaled all of the images to 256px, and divided the data into three sets: train, validation and test set. also some CSV files available for convenience [13]. In this study, just two features (images and labels) are utilized to detect fake image classifiers. Label one represents fake images, whereas label zero represents true images.

IV. SYSTEM STUDY AND ANALYSIS

With the rise of sophisticated deep fake technology, there is a growing concern about its potential misuse, ranging from spreading misinformation to creating fraudulent content. In response to this threat, there is a critical need for robust deep fake detection systems capable of accurately discerning manipulated videos from authentic ones. This project aims to address this challenge by employing a combination of Res Next and LSTM architectures for deep fake detection.

The Existing System uses a variety of methods, such as Recurrent Neural Networks, Adversarial Perturbations, and CrossDomain Fusion, to identify deep fakes. These methods do, however, have several drawbacks that make it more difficult for them to accurately detect edited videos.

Cross-domain fusion methods sometimes have trouble generalizing across different datasets, especially when confronted with changes in camera angles, lighting, or facial expressions. When applied to movies from unexplored domains, their performance may deteriorate, perhaps resulting in false positives or negatives, even though they can attain respectable accuracy inside certain domains. Additionally, Recurrent Neural Networks (RNNs) may have trouble with long-range dependencies and have disappearing or ballooning gradient issues during training, even though they are excellent at capturing temporal dependencies in sequential data. This may make it more difficult for them to recognize minute temporal irregularities that point to deepfake alterations, particularly in videos with intricate and dynamic material. All things considered, the shortcomings of the current system highlight the need for innovative strategies that can overcome these obstacles and improve the precision, resilience, and scalability of deep fake detection systems in practical settings.

Our proposed system for detecting deepfake videos employs a sophisticated blend of advanced deep learning techniques, specifically utilizing LSTM-based neural networks and a pre-trained Res-Next CNN. This combination enables us to effectively analyze sequential temporal patterns within video frames while extracting crucial features unique to each frame. Through extensive training on diverse datasets such as Face Forensics++, Celeb-DF, DFDC, and DeeperForensics-1.0, our model is optimized to handle real-time scenarios with robust performance. Additionally, we have developed a user-friendly front-end application to facilitate seamless interaction, allowing users to upload videos for assessment of authenticity, alongside the confidence level provided by our model. Furthermore, our system incorporates our own dataset, enhancing performance, training accuracy, and enabling real-time identification of fraudulent videos. While acknowledging the potential biases in categorization procedures and the limited scope of existing datasets, our system aims to address these limitations through continuous refinement and adaptation.

V. WORKING FLOW MODEL

This project has been performed in five steps. The broad discussion of these stages is described here. The first stage is choosing the suitable real and fake images dataset from kaggle.com and preprocessing the dataset. Following that, PCA applying for selecting images features after splitting the dataset by using cross-validation (hold-out) (80:20). The next stage is using (CNN) In this section the results of the pre-processing of the images were reviewed, where six images were taken randomly from the dataset that was used in the proposed system. The first three represent the real images and the second three represent the fake ones, where the figure shows the results before and after the pre-processing. The first column represents the original images with RGB color space, while the second column represents the images conversion from RGB into YCbCr color space, the third column represents the images after entering the Gamma correction and finally the fourth. Principal Component Analysis (PCA) is one of the most widely statistical approaches used for feature selection. It has several uses in picture compression, text classification, and face recognition [14]. This is a frequently used method for reducing the dimensionality of a feature collection via a linear transformation. The main goal of PCA is to reduce the original variables to subset of variables by calculating the highest relationship of the original variables [15].

Although the resultant dataset is reduced, but the original data set's features are still retained and removed the redundancy of information [16] and [17]. The number of features in the new dataset may be equal to or lower than that in the original dataset. The principal components are computed by using the covariance matrix. The ability discriminative of the classifiers can be improved by using PCA .

Convolutional Neural Networks is a type of Artificial Neural Network (ANN), also known as Conv-Nets or CNN, That is one of the most effective deep designs for classifying image data and many other applications, including audio recognition, object detection, medical image analysis and natural language processing[19]. The deep architecture of the network produces hierarchical feature extraction, wherein the trained filters of the first layer can be seen as a set of color dots or edges, of the second layer as some forms, the filter of the next layer may learn part of objects and the final layers filters may be able to recognize the objects. [20]. In the proposed method the design is carried out in two stages. This starts first by training the model on the dataset consisting of real images and fake images. As for the second stage which is the testing stage, it is easy to distinguish the test image whether it is fake or real. One of the most common algorithms used in classification is a convolutional neural network (CNN) architecture by created a model from scratch with using six blocks in the CNN architecture. In each block utilize Conv2D with kernel size=5, Max Pooling, batch normalization=64, model = Sequential, dropout, activation = Relu, padding = same, epsilon=0.001, epochs = 100 One iteration on each training data set is represented as an epoch, verbose = 1, shuffle=True and to optimize the network using Adam optimizer = 0.001 and a learning rate 0.0001. Once obtaining the features (components) the features are fed to Convolution Neural Network (CNN) layer of the deep learning model, which further selects the useful features by using their filters. The selected features are supplied to the max-pooling layer to choose the features that have the highest importance value throughout the computation. Gradients will be calculated, and the network's weights will automatically adjust.. We report our detection accuracy in result section. According on the classification results showed that the accuracy of the model with pre-processing is CNN only without PCA and CNN with PCA classifier is 63.86% and 74.26%, respectively. Table 3, Table 4, Table 5, Table 6, Table 7 and Table 8 displays the resulted of confusion matrix with TP, FP, TN and FN values. Our experimental findings without preprocessing stages show that CNN only without PCA achieves 93.16% and CNN with PCA achieves 90.76%. Finally, an additional experiment was conducted by increasing the number of samples entering the CNN network (Training and Testing) which gave the highest accuracy results in image classification, so that the classification accuracy reached 98.04%.

In this section, the results were reviewed for implementing the proposed system and the results were as shown in the explanation below. The research aims to reveal the manipulation that may be present in the images of faces, and therefore it was necessary for us to do a test stating that preserving the image entered into the system for the purpose of testing it and detecting the presence of manipulation in it first.

As a result of the initial processing of the image, it may lead to the burial of some traces of manipulation in the image. So we tested the image by using the technique involved the use of deep learning technology. In this case, we used successive steps for a simple preprocessing (determining the size of the image). We clarified the image and reveal its edges before entering the tamper detection system. We found that any preprocessing of the image, led to undesirable results compared to the results obtained without preprocessing. Where the network of detection of manipulation in faces using the CNN classifier with the preprocessing and the use of canny Detection to detect edges was about 63.86 An additional experiment was used by adding the PCA method after the stage of preprocessing, which produced detection rates of 74.26% for the CNN classifier. This indicates that the use of the PCA as a feature selector has improved the results of detection by a small percentage, but it is not high, because the work of the PCA depends on converting the data of the raw image entered into the detection system into components containing more information that spreads downward from the first component to the last component.

The use of a limited set of component. Since the PCA was used in the form of a feature selector, it relied on the first components in the classification process by CNN, leaving the last components that contain information but are very weak and represent a burden on the classification process. High classification rates and proof that any pre-processing on the image leads to changing or erasing the traces of manipulation encouraged the use of classification techniques approved by CNN, which is CNN with the PCA and without pre-processing, the detection rate of manipulation reached 90.76%, while detection rates using CNN only reached to 93.43. A new attempt was also used by inserting the CNN classifier with PCA on a gray scale image and the results were 86.5%. Finally, CNN classifiers were used to detect forgery and applied to the images directly without using PCA and without any preprocessing, but with more data to see if the network is affected by the number of data that is trained on, and the classification results were 98.04% for the CNN classifier, which is the highest accuracy result in image classification It reached by the proposed system This indicates that the use of PCA in the proposed system affected the accuracy of the results, and that the process of inserting direct images into the CNN classifier has greatly benefited in the process of detecting fraud, as CNN needs a lot of information for the purpose of training and learning from it, regardless of the percentage of information contained in it. Pictures. CNN also made use of the images as more raw data. From the case that the PCA application has, that is, it made more use of the images than the components.

VI. CONCLUSION

I presented a neural network-based approach to classify the video as deep fake or real, along with the confidence of proposed model. Our method is capable of predicting the output by processing 1 second of video (10 frames per second) with a good accuracy. We implemented the model by using pre-trained ResNext CNN model to extract the frame level features and LSTM for temporal sequence processing to spot the changes between the t and t-1 frame. Our model can process the video in the frame sequence of 10,20,40,60,80,100.

The remarkable development of artificial intelligence and with the efficiency of the GAN in generating fake images that are closer to reality, it was necessary to find an efficient way to detect fake images. the final result of our research was to confront the phenomenon of Deepfake offered a detection model for the fake images by utilizing the most popular algorithm in deep learning is (Convolution Neural Network) to detect the fake images. The preprocessing steps start with converting images from RGB to YCbCr color space, after that entering the Gamma correction. finally extract edge detection by entering the Canny filter on them. After that,

utilizing two different method of detection by applying (CNN with PCA) and (CNN without PCA) as a classifiers. The achieved result is better than the listed related work, so utilizing this method enhances the accuracy of classification . From the above results, we conclude the fol lowing:

1. The CNN only is better than a CNN with PCA in the classification accuracy of fake images dataset.
2. CNN is more suitable for large datasets because the network efficiency increases with more data, that is, it gives better results as the training data increases
3. Preprocessing steps when using our dataset give worse results. These steps had a huge impact on decreasing classification accuracy.
4. The type of data used has a significant impact on the categorization accuracy of this work.

REFERENCES

- Guarnera, Luca, Oliver Giudice, and Sebastiano Battiato. "Deepfake detection by analyzing convolutional traces." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 666-667. 2020.
- [2] Wang, L. S. "On the integrated regulation of "deep forgery"." intelligent technology, Oriental Law, 2019.
- [3] Chang, Xu, Jian Wu, Tongfeng Yang, and Guorui Feng. "Deepfake face image detection based on improved VGG convolutional neural network." In 2020 39th chinese control conference (CCC), pp. 7252-7256. IEEE, 2020.
- [4] Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin et al. "{TensorFlow}: a system for {Large-Scale} machine learning." In 12th USENIX symposium on operating systems design and implementation (OSDI 16), pp. 265-283. 2016.
- [5] Chollet, François. "keras. GitHub repository." <https://github.com/fchollet/keras>. Accessed on 25 (2015): 2017. [6] <https://ai.meta.com/blog/deepfake-detection-challenge/>
- [7] M. M. El-Gayar, Mohamed Abouhawwash,,S. S. Askar & Sara Sweidan "A novel approach for detecting deep fake videos using graph neural network"
- [8] F. Song, X. Tan, X. Liu, and S. Chen, "Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients," Pattern Recognition, vol. 47, no. 9, pp. 2825–2838, 2014.
- [9] D. E. King, "Dlib-ml: A machine learning toolkit," JMLR, vol. 10, pp. 1755–1758, 2009.
- [10] N.-T. Do, I.-S. Na, and S.-H. Kim, "Forensics face detection from gans using convolutional neural network," ISITC, vol. 2018, pp. 376–379, 2018.
- [11] X. Xuan, B. Peng, W. Wang, and J. Dong, "On the generalization of gan image forensics," in Chinese conference on biometric recognition. Springer, 2019, pp. 134–141.
- [12] P. Yang, R. Ni, and Y. Zhao, "Recapture image forensics based on laplacian convolutional neural networks," in International Workshop on Digital Watermarking. Springer, 2016, pp. 119–128.