Deep Fake Face Detection using Convolutional Neural Networks

Alben Richards MJ
Department of Electronics Engineering
Madras Institute of Technology
Anna University
Chennai, India
albenmarks@gmail.com

Kaaviya Varshini E
Department of Electronics Engineering
Madras Institute of Technology
Anna University
Chennai, India
kaavyapriya6@gmail.com

Diviya N
Department of Electronics Engineering
Madras Institute of Technology
Anna University
Chennai, India
popdivi25@gmail.com

Prakash P
Department of Electronics Engineering
Madras Institute of Technology
Anna University
Chennai, India
prakashp_mit@annauniv.edu

Kasthuri P
Department of Electronics Engineering
Madras Institute of Technology
Anna University
Chennai, India
kasthurip94@gmail.com

Sasithradevi A
Centre for Advanced Data Science
Vellore Institute of Technology
Chennai, India
sasithradevi.a@vit.ac.in

2023 12th International Conference on Advanced Computing (ICoAC) | 979-8-3503-1821-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICoAC59537.2023.10250107

***Abstract*— More fake face image generators have emerged worldwide owing to the growth of Face Image Modification (FIM) tools like Face2Face and Deepfake, which pose a severe threat to public trust. Although there have been significant advancements in the identification of certain FIM, a reliable false face detector is still lacking. Convolutional Neural Network (CNN) tends to learn picture content representations because of the structure's relative stability. A deep fake face detection model is developed by analyzing the visual features in a face. By the use of deep learning techniques, a CNN model is developed to identify deep fakes.**

Keywords*— Deep fake, Convolutional Neural Network, Fake Face Detection

I. INTRODUCTION

Deep fake face detection is becoming increasingly important today due to the rise of deepfake technology, which is a type of Artificial Intelligence (AI). Deepfake technology is a type of synthetic media that uses convenient methods of machine learning to produce modified videos, photos, or audio that is incredibly realistic and convincing. It creates information that seems genuine but is actually created using deep learning techniques like neural networks, computer vision, and natural language processing. The deepfake technology is becoming more sophisticated, making it challenging to differentiate between real and fake images. Because of its potential for abuse and ethical issues, deepfake technology has garnered attention. In addition to being used for amusement and social commentary, it may also be used for identity theft, digital art, interrupt the privacy of individuals and the manipulation of public opinion or financial fraud. Deepfake technology's fundamental method for producing fake content uses generative models, specifically Generative Adversarial Networks (GANs). A generator and a discriminator are the two neural networks that make up GANs. The discriminator network learns to discern between actual and fake media, whereas the generator network learns to produce fake media by replicating real data. These networks compete against one another in an iterative training process, with the generator aiming to create increasingly realistic material and the discriminator aiming to spot the fakes. Deepfake detection techniques are currently being developed by researchers in an effort to counteract the negative consequences of deepfake technology. These detection methods frequently depend on examining visual artifacts, discrepancies, or statistical anomalies that are peculiar to altered media. Machine learning models learn to identify patterns and characteristics indicating deepfakes.

Various techniques and approaches have been implemented to identify deep fakes and recognize facial expressions. Considering the up-sampling traces present in the deep fakes Zhiqing Guo et al. [1] developed an Adaptive Manipulation Traces Extraction Network (AMTEN), which serves as pre-processing to suppress image content and highlight manipulation traces was proposed. AMTEN exploits an adaptive convolution layer to predict manipulation traces in the image, which are reused in subsequent layers to maximize artifacts by updating weights. Bellahassen Bayar et al. [2] worked on a Deep Learning approach to universal image manipulation detection using a new convolutional layer. This approach automatically learns to detect multiple image manipulations without relying on preprocessing. A new form of convolutional layer that is specifically designed to suppress an image's content and adaptively learn manipulation detection features was implemented. Since deep fakes can be generated by randomly manipulating picture pixels, Richard Zhang et al. [3] worked on making CNN Shift-Invariant. Shift-equivariance is lost in modern deep networks, as commonly used down sampling layers ignore Nyquist sampling and alias: integrate low-pass filtering to anti-alias, a common signal processing technique. The simple modification achieves higher consistency, across architectures and down sampling techniques. Recent research seems to defy the classifier's sensitivity and generalizability issues. These methods distinguish between authentic and artificially produced photos and are applicable to various data generators and datasets. In [4], for instance, the authors trained a common image classifier on pictures created using only one approach (ProGAN) and shown that it can identify pictures created using about a dozen different architectures, training techniques, and datasets that have never been seen before. This classifier is also resistant to scaling, spatial blurring, and JPEG compression.

In the earlier methods employed, finding anomalies in images required statistical analysis and hand-crafted attributes. These methods frequently made use of visual flaws, consistency issues, or statistical irregularities that are common in edited footage. However, their capacity to identify more complex deep fakes is constrained. A deep fake

recognition model with a simple framework but robust detection technique has not yet been developed. In this work, a deep fake face detector based on CNN is developed and trained to identify deep fake images.

II. METHODOLOGY

Deep fake face recognition using a CNN-based approach involves leveraging the power of neural networks to detect manipulated media. The model is trained to distinguish between real and manipulated media. CNNs are particularly well-suited for image and video analysis tasks, making them a popular choice for deep fake detection. The data given as input and passed to the convolutional layers is in image format (pixels in matrix).

A. Deep Fake Face Detection Model

The input image is loaded and pre-processed to standardize its size and format, and to remove any unwanted noise or artifacts that could interfere with subsequent processing steps. A set of features is extracted from the pre-processed image, such as texture, shape, and color information.

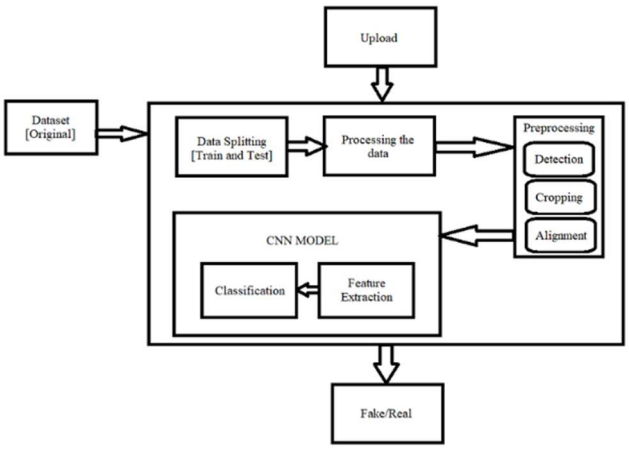


Fig. 1. Pipeline of Deep Fake Face Detection Model

These features are used to capture the characteristics of the image that are most relevant for distinguishing real and fake faces. The feature set is fed into a classification algorithm. Flickr's 140k dataset with 50K real faces and 50K deep fakes was used to train and validate the deep fake face detection model. Fig. 1 illustrates the processes such as data pre-processing, feature extraction and classification for detecting fake faces.

III. SYSTEM ARCHITECTURE

The Deep Fake Face Detection model has five convolutional blocks and a classifier block. There are 13 convolution layers (Conv2D) followed by Pooling Layers, Activation Layers and Dropout Layers. Three Dense Layers and Fully Connected Layers are present in the classifier block. By applying convolutional filters to the input pictures, convolutional layers accomplish feature extraction, collecting regional patterns and structures. To make the feature maps less computationally difficult, pooling layers down samples them. Dropout layer randomly drops out a fraction of neurons in the hidden layers. The final

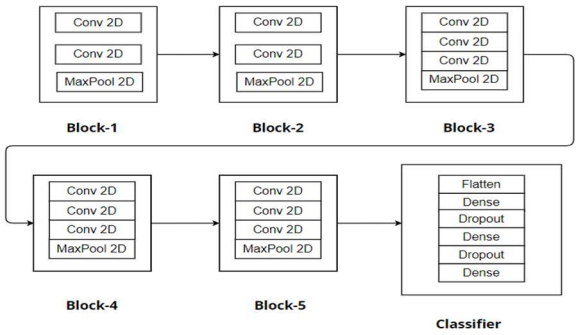


Fig. 2. Architecture of the Deep Fake Face Detection Model using CNN

classification output is given by fully connected layers, which process the retrieved attributes. The CNN model gains non-linearity from activation functions, which enables it to understand intricate connections between input information and output predictions. Fig. 2 shows the architecture of the deep fake face detection model. The input and output are in image format. The preprocessed input data is fed into Block 1 for feature extraction followed by dimensionality reduction and non-linearities introduction. Lower-level blocks detect simple and local features, such as edges and textures, while higher-level blocks combine these features to detect more complex patterns and objects. Each block captures more abstract and high-level features as we move deeper into the network.

IV. SYSTEM IMPLEMENTATION

A. Data Preprocessing

Pre-processing can be used to boost system performance before the feature extraction process. A variety of processes are involved in pre-processing a picture, such as face alignment and identification, lighting, posture, occlusion, and data augmentation correction. Images that are realistic include different sizes, stances, zooms, lighting, noise, etc. The Data Augmentation approach is employed to make the network resilient to these often-occurring effects. The network will experience these effects during training by rotating input pictures at various angles, flipping images along various axes, translating/cropping or padding the images. To simplify the complicated pixel values, images will be downsized with three red, green, and blue channels (three planes) to grayscale with just one channel (single plane).

B. Feature Extraction

After the image data has been pre-processed, it is transferred to the CNN model for feature extraction, where convolution is carried out on the 2D array of pixels with the aid of filters, leading the Feature Map. Mathematically convolution is given by:

$$O(x,y) = \sum_{k=1}^{a} \sum_{l=1}^{b} I(x+k-1, y+l-1) K(k,l) \quad (1)$$

where x ranges from 1 to A - a + 1 and y ranges from 1 to B-b+1. Equation 1 describes the outcome of the convolution process and refers to the Feature Map, which provides details about the picture's borders, corners, and other features. Then,

additional layers will use this Feature Map to look for more features in the input picture. The feature map is made non-linear with the use of the ReLU activation function. ReLU is straightforward to calculate and provides an error-backpropagation gradient that is predictable. Max Pooling is then applied to the resulting matrix. The CNN model maintains the same picture size throughout.

This is achieved by "same" padding where zeros are padded on the empty arrays so that the size remains same and only the spatial dimension(density) of the matrix reduces layer by layer. Some neurons associated with the Dropout Layer go inactive at random. The SoftMax activation function will receive the output data as a parameter. The fully connected layer uses this function to transform a vector of numbers into a vector of probabilities. In Deep Fake Face Detection Model "1" indicates that the input image is real whereas "0" indicates that the input image is deep fake.

V. Results And Discussion

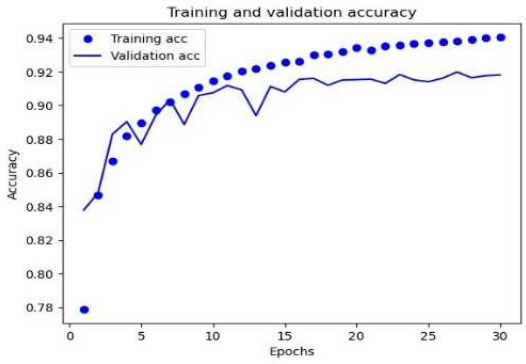
A. Evaluation Parameters

The parameters used to assess the model performance include accuracy, loss, and confusion matrix. Values for the four combinations of true and anticipated values—True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are provided by the confusion matrix. TP, FP, TN, and FN are used to calculate the evaluation metrics. Forecasts for correct predictions are denoted by TP, those for incorrect predictions by FP, those for correct inaccurate predictions by TN, and those for wrong incorrect predictions by FN. Accuracy is the fraction of predictions the model got right. In cross-entropy loss function each predicted class probability is compared to the desired output calculated. A perfect model has a cross-entropy loss of zero. Cross-entropy is given by equation 2.

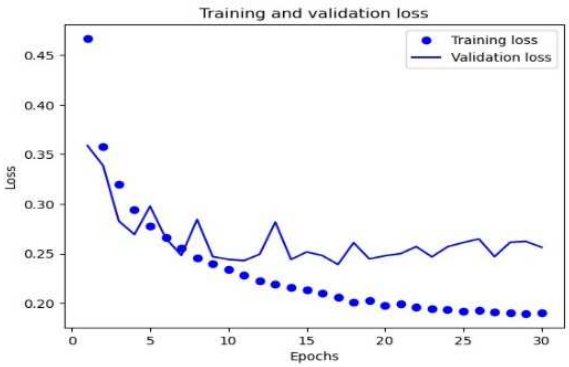
$$L_{CE} = -\sum_{i=1}^{c} t_i \log(p_i) \quad (2)$$

where t_i is the truth label, c is the number of classes(expressions) and p_i is the softmax probability value for the i^{th} class. Precision is a performance metric used in machine learning and statistical analysis to measure the proportion of TP (correctly identified positives) out of all predicted positives. F1 score is calculated as the weighted average of precision and recall. Recall is calculated as the ratio of TP to the sum of TP and FN. The F1 score with a higher value indicates better performance.

B. Results of Deep Fake Face Detection Model



(a)



(b)

Fig. 3. Accuracy And Loss for Deep Fake Face Detection Model

The Deep Fake Face Detection Model is trained for 30 epochs after which the respective accuracy and loss curves were plotted after training and validation. From Fig. 3, it can be observed that the training accuracy and validation loss are greater than validation accuracy and training loss respectively. Accuracy is expressed in terms of percentage value. Loss is calculated using binary cross entropy function which has value in the range of zero to positive infinity. It is a measure of the cross-entropy between the true label and the predicted label. The result of loss function is a scalar value representing the overall performance of the model on the classification task.

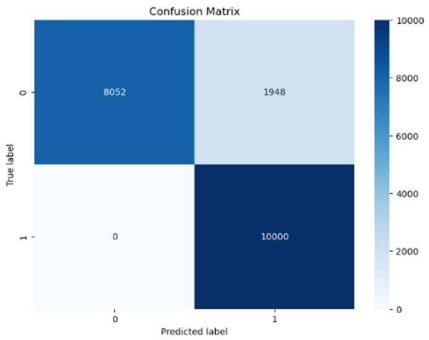


Fig. 4. Confusion Matrix for Deep Fake Face Detection Model

The classification is binary, in which "0" means "fake" and "1" means "real". It is inferred from Fig. 4 that the TP and TN values are higher than the FP and FN values. This shows that the model performance is effective. The results and discussion of a deep fake face detection model involve, evaluating the performance of the model on a set of test data, and comparing its performance. From Fig. 4, the total number of TP is 8052 and total number of FP is 1948. Precision is calculated as 81% and the F1 score is 91% for the Deep Fake Face Detection Model. Real images and Fake images generated from StyleGAN and other similar architectures has been used as input to the Deep Fake Face Detection Model. The anomaly in the fake images caused by up-sampling techniques has been learnt by the model during the training phase.

Table 1 shows that the Deep Fake Face Detection Model was able to predict deepfakes and real faces based on the input having an efficient validation accuracy, validation loss, training loss, and training accuracy. Dropout Layers have been used to prevent overfitting of the model. Hence the model performs effectively with a minimal divergence in accuracy.

TABLE I. Accuracy and losses for training and validation

Model	Accuracy (%)	Loss
Training	94.08	0.1902
Validation	91.82	0.2565

The model has demonstrated good generalization capabilities, successfully detecting deep fakes across diverse test cases of varying image qualities.

(a) (b)

Fig. 5. Detection of Fake Face

(a) (b)

Fig. 6. Detection of Real Face

Fig. 5 shows the model prediction for deep fakes based on the pattern features and anomaly in the image. Fig. 6 shows the model prediction for real faces.

VI. Conclusion And Future Scope

A novel Deep Learning approach is developed to identify deep fake faces in images. CNNs automatically learn hierarchical representations of features from raw image pixels through convolutional layers with shares weights, providing translational invariance and robust pattern recognition. Despite other advanced algorithms like SVMs, Random Forests and Gradient Boosting, which require extensive feature engineering and may not handle complex image data as efficiently, CNNs remain highly effective. The suggested approach trains a neural network to discriminate between real and fake faces using deep learning techniques and an extensive dataset of actual and fake face images. The system was tested using a number of metrics, and the results demonstrated that it is more effective at identifying deep fakes. The evaluation outcomes implied that, even when endured with fraudulent faces, the suggested system can recognize faces with high accuracy of 91.82% and a F1 score of 91% which has been developed using an easily understandable CNN network rather than complex techniques like SVM, Boosting, Transfer Learning and other advanced architectures.

There are numerous applications in various fields such as security, entertainment, forensics, journalism, healthcare, sports, marketing and interactive platforms. This project broadens the forensics department's perspectives and ideas. Facial Manipulations such as Identity Swap, Expression Swap, Attribute Manipulation and Entire Face Synthesis in security systems can be recognized. The future scope for deep fake face detection involves a multidisciplinary approach, incorporating advancements in computer vision, machine learning, and psychology to develop more robust, accurate, and reliable detection techniques. Deep Fakes present a number of obstacles, and developing effective detection and mitigation techniques will need continued research collaboration and cooperation between academia, business, and policymakers. Deep Fake recognition algorithms can detect the subtle inconsistencies in images or videos that are characteristics of deepfakes, while face recognition algorithms can identify and verify an individual's identity based on facial features. The use of these algorithms can help prevent the spread of fake news, protect against identity theft and fraud, and enhance the accuracy of medical diagnosis and treatment plans.

References

[1] Zhiqing Guo, Gaobo Yang, Jiyou Chen, Xingming Sun (2021) “Fake face detection via adaptive manipulation traces extraction network” in Computer Vision and Image Understanding-Volume 204.

[2] Belhassen Bayar and Matthew C. Stamm (2016) “A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer” in IH & MM Sec '16: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security.

[3] Richard Zhang et al., (2018), “Making Convolutional Neural Networks Shift-Invariant Again” in ICML 2019.

[4] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot...for now. In IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[5] Carlini, N. and Farid, H. (2020) “Evading deep-fake-image detectors with white-and black-box attacks” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[6] L. Tran, X. Yin, and X. Liu, “Representation learning by rotating your faces,” IEEE transactions on pattern analysis and machine intelligence (TPAMI), vol. 41, pp. 3007–3021, 2018.

[7] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, “StarGAN v2: Diverse image synthesis for multiple domains,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8188–8197.

[8] Zhou, P., Han, X., Morariu, V. I., and Davis, L. S. Learning rich features for image manipulation detection. In IEEE International Conference on Computer Vision (ICCV), 2018.

[9] Frank, Joel & Eisenhofer, Thorsten & Schönherr, Lea & Fischer, Asja & Kolossa, Dorothea & Holz, Thorsten. (2020). Leveraging Frequency Analysis for Deep Fake Image Recognition.

[10] McCloskey, S. and Albright, M. Detecting GAN generated imagery using color cues. arXiv preprint arXiv:1812.08247, 2018.

[11] Marra, F., Gragnaniello, D., Verdoliva, L., and Poggi, G. Do GANs leave artificial fingerprints? In IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019.

[12] Yu, N., Davis, L. S., and Fritz, M. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In IEEE International Conference on Computer Vision (ICCV), 2019.