Detection and Generation Analysis with Generative Adversarial Networks and Autoencoders

Prof. Rajni Jindal[1], Gaurav Singh[2], Himanshu[3], Himanshu Sherwan[4]
[1,2,3,4]*Department of Computer Engineering, Delhi Technological University, Delhi, India*
[1]rajnijindal@dce.ac.in, [2]gauravpksingh@gmail.com,
[3]jorwalh@gmail.com, [4]himanshusherwan00@gmail.com

***Abstract*—Deep Learning techniques have solved many of the challenges inherent in Enterprise domains including Big Data Analytics, Computer Vision and Human-level Control. Of all the deep learning advancements the most recent application utilized in deepfake. Deepfakes have since increased the documentation of fake material, ever since deep learning technologies emerged. The Deepfake works on the basis that the face of one person can be replaced with the face of another in an image or video, which may pose a security risk. Deep fakes are useful in entertainment, creativity, digital effects, and learning and can be used to claim, the general use of deep fakes can cause widespread harm to the illustration, Recently, there have been more women who are spreading false information and producing wholly false pornographic content. Deepfakes: Building the most realistic and accurate deepfakes takes time, effort, and computing power, however, it's relatively easy to create good quality deepfakes and detecting these deepfakes because they are so easily created, is another story, so that creates a need to pursue the methods to detect deepfakes. This paper analyses deepfake generation with GANs and Autoencoders measure the productivity of each technique in producing a deepfake using the aforementioned methods and deepfake detection using DenseNet and ResNet and then modelled a deepfake detection by merging the feature engineering architectures of both DenseNet and ResNet to measure the effectiveness in deepfake detection.**

***Index Terms*—deepfake generation, GANs, Autoencoders, deepfake detection, DenseNet, ResNet, Convolutional Neural Networks**

I. Introduction

The face is one of the most unique things about humans. The name "deepfakes" is a combination of "Deep Learning" and "Fake". Using numerous deep learning algorithms, it is possible for an individual's faces to be replaced by another individual's face in a sufficiently realistic way. Deep fakes is a budding field of intelligence wherein the face of one person is laid over the structure of another person's face. On a larger scale, deepfakes are synthetic content and can be further divided into three broader categories, i.e., face-swap, puppet-master, and lip-sync. Faceswap is the replacement of a face in a video, with another face of some other person. Lip-sync, in contrast, is a term used for the method of editing a source video such that the mouth region is synchronized with an audio recording, while in puppet-master the legitimate method a target person is animated (head shifts, eye movements, facial expressions) through a performing artist before a camera, who acts out what the performer needs their puppet to say and do [2].

This phenomenon began in 2017, where a Reddit user applied deep learning methods to replace an individual's face in a pornographic video, creating very realistic forged content [3] [4]. The use of technology, especially in mobile phone, camera devices and advance artificial intelligence tools, applications, etc. have led to the production of massive amount of doctored videos and images, of several celebrity and politicians just to defame, spread false news for personal interests, or harass other people [5]. Falling back on a treasure of videos and pictures has been key. Deepfake models require extensive training, and, as such, high-quality deepfakes do require large sets of data. Such data was difficult to collect in the past, but with the internet revolution, it has become possible today. Moreover, high capacity computer systems are now easily accessible and allow to train such a neural network much easier, which have leaded to the exponential increase of deepfake content.

In the past, deepfakes of politicians including Barack Obama, Joe Biden, and Donald Trump have made the rounds. In 2018, an altered video was generated where former US president Barack Obama is shown saying statements he has never said. In February 2020, Manoj Tiwari, the Delhi BJP president, created a bilingual election appeal for votes in Haryanvi and English [4] using deepfaking techniques. Such malicious application of deepfakes is an extremely dangerous threat to society and it may lead to the dissemination of false information, particularly, on social media. Despite all of the problems facing the Former

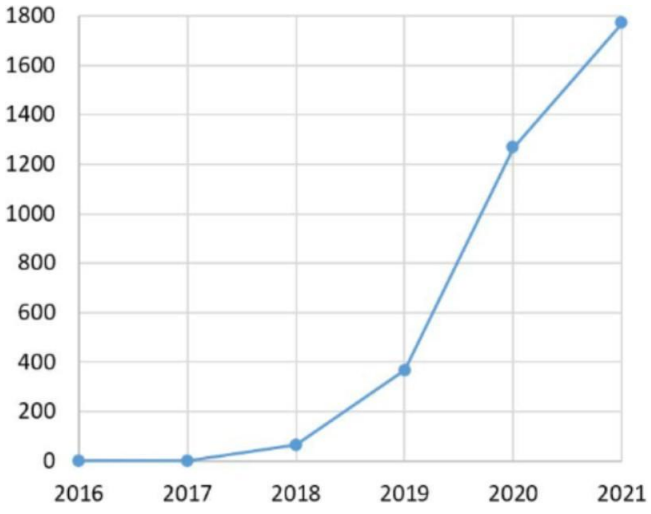

Fig. 1. Papers related to deepfakes in years from 2016 to 2021

generation of deepfakes, there are several positive applications as well; in the entertainment industry, it is used for example to swap the face of an actor with the face of an actor in the background and also as a means for the detection of anomalies in X- ray images, Generative Adversarial Networks GANs are used and for the educational purpose.

Although some deep fakes may be created by means of, traditional visual effects or computer graphics techniques, the so-called modern way of creating deep fake is through deep learning models, such as Autoencoders and Generative Adversarial Networks GANs [5]. These referred models are used to analyze the facial expressions and movements of an individual, which make it feasible to synthesize facial images that reflect similar expressions/movements of another person. The majority of tech titans are seeking out solutions for detecting fake content, in order to stem the tidal wave of fake content that has begun flooding the internet. Recently, many established corporations like Facebook, Apple, Google and Microsoft Corp. are researching on deepfake detection to foster much more research in the field of deepfake. The fierce interest from companies like Google, Meta and Microsoft Corp. highlights the importance of the issue. The Fig. 1 adapted from [6] shows a substantial and statistically significant increase in the volume of academic papers related to deep fake technology in the last few years. The paper depicts the potential ways of generating deepfakes using deep-learning-based techniques like autoencoders and GANs and we suggest to use the merged version of two different CNN architectures over the base CNN model being used for detection, to join these two models we will use ensemble learning as well as transfer learning.

II. LITERATURE REVIEW

A. Deepfake generation methods

Goodfellow, et al. [11] introduced a new approach to generative model estimation via an adversarial method.

We present adversarial network which trains two models (generative model G which learns the data distribution and generative model D that discriminates whether a sample was born out of training data or the generators output,I.e, the likelihood that a sample came from the training set rather than from G. This way both models learn and improve to the point where the image generated by the generator model can no longer be found away from real input data. The adversarial networks were trained using the MNIST, Toronto Face Database (TFD) and CIFAR-10 datasets [11]. Liu and Tuzel [12] introduced Coupled Generative Adversarial Networks (CoGAN), which is an extension of the Generative adversarial networks (GANs), that allows us to learn joint distributions of multi-domain images in two different domains. Coupled Generative Adversarial Networks (CoGAN) consists of two GANs: G1 and G2, both of which independently perform image synthesis in a single domain. Experimental with MNIST digits and image faces with NYU dataset and RGBD dataset were performed on CoGAN which showed that how it could be capable of generating corresponding images even if there is no special training for the same to work with [12].

It's been a trend over the last few years to use generative deep neural networks to manipulate faces. Karras et al. [13] proposed a generative adversarial network known as styleGAN Zhu et al. proposed Cycle- GAN a Generative Adversarial Network based face-swapping method. [14]. Also, for swapping the face of single individual in an image or film with the face of different individual tools like DeepFakes [15] and FaceSwap [16]can be applied.

B. Deepfake detection methods

Deepfake images detection and deepfake videos detection are two different things. Deepfake image detection utilizes the image pixels [17], and analyzes the noise level [18] to perform manipulation/AI generation detection in the image. The references in [19] classify deepfake detection methods into holistic-based and feature-based matching. Holistic methods, such as PCA, SVM [21], their main goal is to minimize data dimensionality creating a set of linear combinations of image pixels which are provided to a binary classifier. This method identifies localized features of deep fake images such as (eyes, mouth, nose) [20]. Deep Learning Techniques based on CNNs The Most Successful Technologies Used individually in Deep fake identification and detection[22]. Some classic CNN architectures are VGG16 [24], XceptionNet [25], InCeptionV3 [23].

The Combination of CNN and LSTM Architecture [26] Where CNN assists in closure or opening of eye and LSTM assists in discovering temporal information in since passage of eye is a decent relationship of close-by frames. In paper [27], it proposes a method for segmentation detection

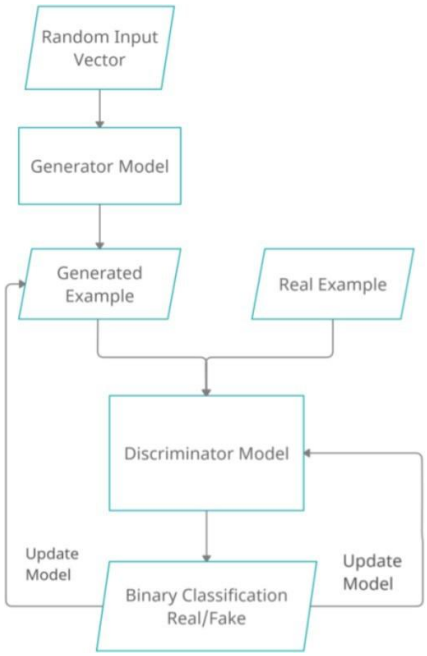

Fig. 2. Generative Adversarial Network Architecture

deep fakes which are created by means of overlaying the goal character's face on the supply face while detection is executed through estimating the face key factor distinction after which feeding to the SVM for the final selection.

III. PROPOSED METHODOLOGIES

A. Goals and Objectives

This paper focuses on exploring and analyzing deep fake generation based on GANs and Autoencoders and further explores the feasibility of implementing integrated ResNet50 and DesNet121 architectures for deep fake detection.

B. Dataset

Kaggle data set of real and fake images is used for detection of deepfakes; This dataset is composed of a total of 190,000 images divided into three sets: train, test, and validation set.

C. Deepfake using Generative Adversarial Network

GANs or Generative Adversarial Network is a model that uses deep learning like neural networks. The premixes competition in the network, GAN model architecture optimize it by utilizing two sub-models i.e, generator and discriminator. Generator tries to generate new instance of image from latent vector given to it and discriminator figure out whether instance generated is comes from domain provided or fake, which was generated by generator. The second step is to train the network where the generator will be trained to produce images that are as real as possible so that they are not detected by the discriminator. The objective is to generate image that the discriminator will misclassify [9] - [10].

In the case of GANs, the generator model inputs a gaussian distributed fixed-length random vector to generate sam- ples in the target domain. This multidimensional vector space is known as latent space, and it corresponds to a set of points in the domain which acts as a compressed representation of the data distribution during training. The generator maps latent space points to meaningful feature vectors, so sampling points from latent space produces different output examples by generating new input points from the latent space [10]. Discriminator is the one that takes image as input and try to classify it as real or fake, in this aspect it works similar to any other neural network. The convolutional neural network produces a single value per image [9].

D. Deepfake using Autoencoders

Autoencoders, one of the most important unsupervised learning algorithms, can be used to compress the data and reduce its dimensionality. Autoencoders are a type of neural network that can learn to reconstruct images, text and other data from compressed versions of themselves. Now first the original image is passed through the encoder to create a lower dimensional representation of that same face, called as a base vector for that face also as latent face. The latent face reconstructed when the latent face is passed through the decoder. Since Autoencoders are lossy [8], the reconstructed face will lack exactly the same information than was there before. In generating a deepfake, training phase A, B in which decoder A is not trained on images of B and vice versa. But all the base vector is generated using same encoder. Upon completion of the training process, A is passed to Decoder B, and the base vector from B is passed to Decoder A; if the network has generalised sufficiently, the latent space or base vector will reflect facial expressions and orientations, resulting B visage generation with A expressions and orientations and vice versa [8]. When we train the autoencoder, it is given a set of images and the encoder-decoder layer will adjust parameters to produce an output similarly to the input images.

E. Image detection model

We evaluate and analyze a method for the classification of both real and fake face images using hybrid convolutional neural networks (CNNs) which are ResNet-50 and DenseNet- 121 architectures in this research paper. As part of our preprocessing, we have implemented the data prepro- cessing and augmentation strategies, such as image resizing, random horizontal flipping, rotation, and color jittering, and afterwards we used it with the training dataset to improve the model. In order to ensure homogeneity and regularity during model training, we also implement normalization methods over all the datasets. The hybrid model architecture employed features from ResNet50 and DenseNet-121 networks which are later concatenated through fully connected layers with rectified linear unit (ReLU) activation functions and 30% dropout regularization to evade the risk of overfitting. The model is trained using the Adam optimizer with the learning rate equal to 0.0001 and weight decay equal to 0.001 to avoid model overfitting. Moreover, we employ the learning rate scheduler based on ReduceLROnPlateau to dynamically modify the learning rate while training the data, resulting in the model being more optimized.

IV. RESULT AND DISCUSSION

Insights from deep learning architectures for image generation The styleSwin achieved significantly better performance as compared to StyleGAN, with an FID lower with respect to 3.25 [30]. VAEBM performed average, it score FID scores of 20.38 on CelebHQ and 5.31 on CelebA 64 dataset, indicating some room for further optimization [31]. FaderNet has a reasonable accuracy of 88.6% on CelebA, revealing how good it is at preserving the invariance on the attributes [32]. Importance of exploring and refining deep learning architectures for image generation tasks is presented by these findings.

We utilized CNN-based models for classifying the fake and real images. The initial resnet-50 model produced an accuracy of 75.24%, with precision and recall scores equal to (0.80, 0.68) respectively and the F1 score for the test data set equal to 0.73. The second DenseNet-121 based model achieved a test accuracy of 79.13% with a precision of 0.77, recall of 0.71 and test F1 score of 0.74. Both models have been fine-tuned to get the best results.

The Hybrid model after 20 epochs obtained a final validation accuracy of 85.4% on both training and validation sets. This results in an test-accuracy of 83.7%, average-

precision of 0.81, average-recall of 0.84, F1 score of 0.82 and a confusion matrix that indicates the successful separation of real and fake faces by the model. Training and validation epoch-wise visualizations show the convergence behaviour of our model, resulting in a good facial classifier, indicating the usefulness of hybrid CNN architectures on facial images.

TABLE I
PERFORMANCE METRICS OF CNN MODELS

Model	Test Accuracy	Precision	Recall	F1 Score
ResNet-50	75.24%	0.80	0.68	0.73
DenseNet-121	79.13%	0.77	0.71	0.74
Hybrid Model	83.7%	0.81	0.84	0.82

V. CONCLUSION

This study presents hybrid CNN model for identifying real and fake faces as final conclusion. Combining the ResNet-50 and DenseNet-121 architecture, while applying data preprocessing and transformation, we achieved test accuracy of 83.7% and validation accuracy of 85.4%. We can do this by concatenating the features obtained from both the models. The model is demonstrating classification capabilities as you can see precision, recall and F1 score metrics are all doing great. The hybrid CNN has demonstrated its great potential on facial image analysis tasks based on our findings.

REFERENCES

[1] Zhang, Jixin, et al. "A heterogeneous feature ensemble learning based deepfake detection method." ICC 2022-IEEE International Conference on Communications. IEEE, 2022.

[2] Agarwal, Shruti, et al. "Protecting World Leaders Against Deep Fakes." CVPR workshops. Vol. 1. 2019.

[3] O¨ hman, Carl. "Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography." Ethics and Information Technology 22.2 (2020): 133-140.

[4] Garg, Diya, and Rupali Gill. "Deepfake Generation and Detection-An Exploratory Study." 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). Vol. 10. IEEE, 2023.

[5] Nguyen, Thanh Thi, et al. "Deep learning for deepfakes creation and detection: A survey." Computer Vision and Image Understanding 223 (2022): 103525.

[6] https://app.dimensions.ai

[7] Corcoran, M., and M. Henry. "The Tom Cruise deepfake that set off 'terror' in the heart of Washington DC." ABC News. Retrieved 13 (2021).

[8] https://www.alanzucconi.com/2018/03/14/understanding-the- technology-behind-deepfakes

[9] Das, Manoj Kumar, et al. "Deepfake Creation Using Gans and Autoencoder and Deepfake detection." 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN). IEEE, 2023.

[10] Sontakke, Nikhil, et al. "Comparative Analysis of Deep-Fake Algo- rithms." arXiv preprint arXiv:2309.03295 (2023).

[11] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

[12] Liu, Ming-Yu, and Oncel Tuzel. "Coupled generative adversarial networks." Advances in neural information processing systems 29 (2016).

[13] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[14] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.

[15] https://github.com/MarekKowalski/FaceSwap

[16] https://github.com/deepfakes/faceswap

[17] Li, Yuezun, and Siwei Lyu. "Exposing deepfake videos by detecting face warping artifacts." arXiv preprint arXiv:1811.00656 (2018).

[18] Li, Lingzhi, et al. "Face x-ray for more general face forgery detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[19] Zhao, Wenyi, et al. "Face recognition: A literature survey." ACM computing surveys (CSUR) 35.4 (2003): 399-458.

[20] Silva, Samuel Henrique, et al. "Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models." Forensic Science International: Synergy 4 (2022): 100217.

[21] H. Agarwal, A. Singh, and D. Rajeswari, "Deepfake Detection Using SVM," Proc. 2nd Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2021, pp. 1245–1249, 2021, doi: 10.1109/ICESC51422.2021.9532627.

[22] Jolly, Vedant, et al. "CNN based deep learning model for deepfake detection." 2022 2nd Asian conference on innovation in technology (ASIANCON). IEEE, 2022.

[23] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[24] Simonyan, Karen, and Andrew Zisserman. "Very deep convolu- tional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[25] Chollet, Franc¸ois. "Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[26] Chen, Joy Iong Zong, and S. Smys. "Social multimedia security and suspicious activity detection in SDN using hybrid deep learning technique." Journal of Information Technology 2.02 (2020): 108-115.

[27] Yang, Xin, Yuezun Li, and Siwei Lyu. "Exposing deep fakes using inconsistent head poses." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[28] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[29] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[30] Zhang, Bowen, et al. "Styleswin: Transformer-based gan for high- resolution image generation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[31] Xiao, Zhisheng, et al. "Vaebm: A symbiosis between variational autoencoders and energy-based models." arXiv preprint arXiv:2010.00654 (2020).

[32] Lample, Guillaume, et al. "Fader networks: Manipulating images by sliding attributes." Advances in neural information processing systems 30 (2017).