A Review on Deepfake Multimedia Data on Heterogeneous Filter Effects

Shruthi.S[1], Manjunath R[2]

[1]*Assistant Professor, Department of Computer Science & Engineering, RR Institute of Technology, Karnataka, India*

[2]*Professor & Head , Department of Computer Science & Engineering, RR Institute of Technology, Karnataka, India*

ABSTRACT

Deep learning is being used so frequently now that media synthesis and manipulation have never been more realistic. The go-to tool for controlling the media is now Deepfake. Despite having uses in the entertainment industry, this technology is also susceptible to political manipulation and other issues. This technology has made significant advancements and supports a variety of applications in TV channels, the video game industry, and the film industry, such as improving visual effects in movies, as well as a number of illegal actions, like spreading false information by imitating well-known individuals. Research on DeepFake identification utilising deep neural networks (DNNs) has drawn more attention in order to recognise and categorise DeepFakes. DeepFake is essentially regenerated material that has had some information added to it or replaced by using the DNN model

KEYWORDS: *Deep learning, DeepFake, CNNs, GANs, image classification, image forensics.*

1.INTRODUCTION

Deep generative models (DGMs), particularly variational auto encoders [1] and Generative Adversarial Networks [2], have recently advanced to new degrees of realism in media synthesis and manipulation. Medical imaging, digital forensics, and art production have all been touched by DGMs. The emergence of 'deepfake,' an infamous technology that leverages DGMs to superimpose facial photos of a target person over those of a source person, has revealed the dark side of DGMs. The effectiveness of deep learning models can no longer be ignored; in fact, they are steadily replacing most technology and are being quickly adopted by numerous research communities and huge IT corporations. The recent growth of digital data throughout the Internet, as well as its relevance in everyday life, such as digital marketing, legal forensics imagery, medical imagery, sensitive satellite image processing, and many other applications, cannot be overlooked. Moreover, digital data in different applications are evolving in such a way that they are also fueling an uptick in cybercrime. In the Figure 1, the trend indicates serious vulnerabilities and a decrease in the trustworthiness of digital data. Moreover, discerning whether the acquired digital data are authentic or altered and legitimizing digital documents are currently major problems.

Figure 1. Deepfake creation with frame of Celebs from left to right which creates the target image

When early picture-based algorithms for deepfake image detection focused on prominent art facts, it was discovered that, when compared to convolutional neural networks, these methods do not generalise as well to samples originated from unknown generators with latent arte facts (CNNs). The core component of CNNs, Adaptive Weighted Filters (AWFs), has demonstrated their superiority in the field of pattern recognition. In other words, the model is trained on DeepFake datasets and then put to the test in trials to assess how well it works. In this essay, we'll go over the various DeepFake detection approaches for images and videos. We'll also go over the DeepFake detection algorithms and datasets. Studies on DeepFake production and detection in images, sounds, and videos have recently been published. Deep learning and computer vision techniques, such as GANs [3] and autoencoders [4], have allowed for the creation of superre alistic false images and movies known as DeepFakes. DeepFakes (a mix of the phrases "deep learning" and "fakes") allow attackers or even non-technical machine learning users to alter a picture or video by swapping out the material and generating a new image or video that humans and computers cannot distinguish. People's faith in digital media content has been eroded as a result of the advent of DeepFakes, as they can no longer accept the visuals they are seeing. Research on identifying or detecting fraudulent modified media is considered standard research in the absence of deep learning. Currently, generative deep models are quite effective at building DeepFakes that are difficult to differentiate using traditional approaches.

2.TECHNICAL BACKGROUND

2.1 CNN Background

Convolutional, pooling, and fully connected layers are the three types of layers that make up the CNN model's basic structure. The convolution layer is responsible for feature extraction.

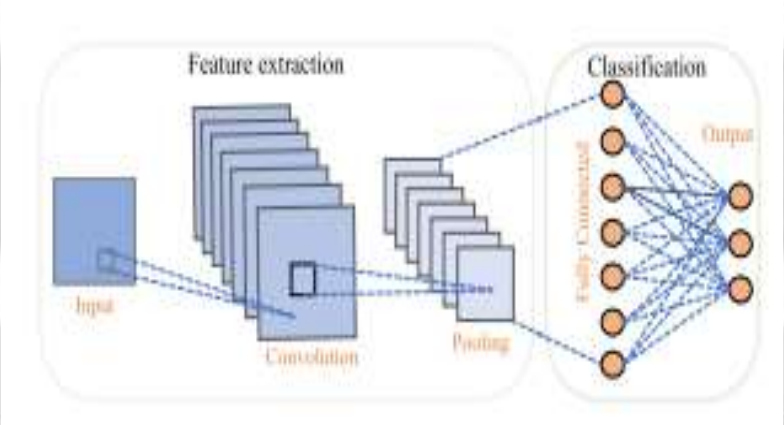

Figure 2. Architecture of CNN

The feature map is constructed by applying an array of numbers (kernel) across inputs (tensor) in the convolutional process. The element wise product between each element of the kernel and the input tensor is used to generate a feature map, and the outputs are summed to get the element of the kernel. To create the elements of the feature map for that kernel, the kernel convolves across all the elements on the input tensor is describe in Figure 2. Implementing the convolution technique with different kernels can provide an arbitrary number of feature maps. Forward propagation occurs during training, while backpropagation occurs when the gradient descent optimization algorithm changes the learnable parameters (kernels and weights) based on the loss value. The feature value (Z l i,j,k) at location (i, j) in the k th feature map of the l th layer in [5] is as follows:

$$Z^l_{i,j,k} = (W^l_k)^T x^l_{i,j} + b^l_k \qquad (1)$$

where Wl k and b l k are the weight vector and bias term of the k th filter of the l th layer, respectively. x l i,j is the input patch centered at location (i, j) of the l th layer. Then, a nonlinear activation function is applied to detect nonlinear features such as sigmoid, tanh and ReLU[6]. A nonlinear activation function A(·) can be expressed as:

$$a^l_{i,j,k} = A(Z^l_{i,j}), \qquad (2)$$

where a l i,j,k is the output value after applying the nonlinear activation function. A pooling layer provides a typical down sampling operation to reduce the dimensionality of the feature maps to introduce translation invariance to small shifts and distortions and thereby decrease the number of subsequent learnable parameters. The pooling function is pool(·); for each feature map a l :,:,k , we have:

$$y^l_{i,j,k} = pool(a^l_{m,n,k}), \quad \forall (m,n) \in R_{i,j}, \qquad (3)$$

where Ri,j is a small neighborhood in the vicinity of the place (i, j). The CNN's ultimate outputs, such as the probabilities for each class in classification tasks, are completely linked layers. In the final fully connected layer, the number of output nodes is usually equal to the number of classes. Each fully linked layer is followed by a nonlinear function, such as ReLU. Finally, a loss function is computed to determine whether the CNN's forward propagation output predictions are compatible with the provided ground truth labels. CNN's loss can be computed using the following formula:

$$\mathfrak{L} = \frac{1}{N}\sum_{n=1}^{N} \ell(\theta; y^{(n)}, o^{(n)}), \qquad (4)$$

where the number of input-output relations (x (n), y (n) is denoted by N, where x (n) is the nth input data, y (n) is the target label, and o (n) is the CNN output [13]. By minimising the loss function, training a CNN identifies the global minima, which indicate the best-fitting collection of parameters. Many CNN models, such as AlexNet, are currently available

.

2.2 RNN Background

An RNN is a neural network in which the output from the previous step is used as input in the next phase. All inputs and outputs in typical neural networks are independent of one another; however, in some situations, such as when predicting the next word of a phrase, the prior words are necessary, and therefore, the previous words must be remembered. Consequently, RNNs were created, which use a hidden layer to overcome the problem. The hidden state, which remembers certain information about a sequence, is the most significant aspect of RNNs. RNNs have a ‘‘memory’’ that stores all information about the calculations. This memory utilizes the same settings for each input since it produces the same outcome by performing the same job on all inputs or hidden layers. Unlike in other neural networks, this method minimizes the complexity of the parameters.

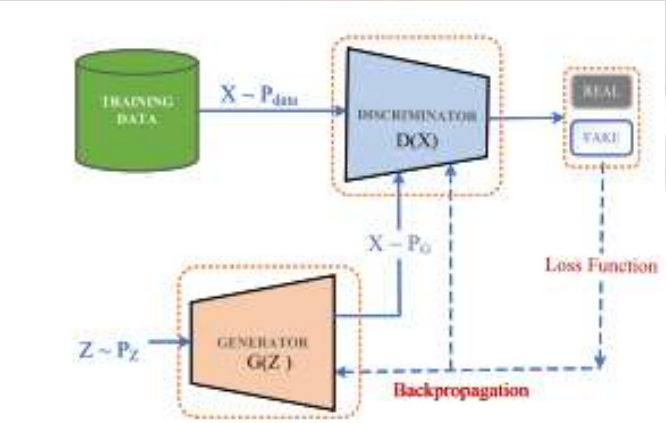

Figure 3. Architecture of a GAN.

The recurrent layers, also known as hidden layers in RNNs, are made up of recurrent cells whose states are influenced by both previous states and current input via feedback connections. The classic recurrent sigma cell and LSTM with only input and output gates are depicted in Figure 3. The LSTM mathematical expressions are as follows:

$$
\begin{aligned}
i_t &= \sigma(W_i h_{t-1} + W_i x_t + b_i) \\
\hat{c}_t &= tanh(W_{\hat{c}} h_{t-1} + W_{\hat{c}} x_t + b_{\hat{c}}) \\
c_t &= c_{t-1} + i_t \cdot \hat{c}_t \\
o_t &= \sigma(W_o h_{t-1} + W_o x_t + b_o) \\
h_t &= o_t \cdot tanh(c_t), \qquad (5)
\end{aligned}
$$

where xt , ct , ot and ht denote the input, the recurrent information, and the output of the cell at time t, respectively; Wi , Wcˆ , and Wo are the weights; and b is the bias. ct denotes the cell state of LSTM, and the operator '·' denotes the point wise multiplication of two vectors.

2.3 GANS Background

GANs are a revolutionary tool used for teaching generative models to generate realistic examples from a data distribution [2]. Basically, GANs are a combination of two neural networks: the generator, (G), and the discriminator, (D). These two neural networks compete in a dynamic minimax game. The intuition behind this idea is that G attempts to create fake samples, while D attempts to determine which samples are fake and which are real. If the two models are allowed to compete for a long time, they will ultimately improve. In other words, the generator G aims to capture the data distribution, whereas a D aims to estimate the probability that a sample comes from the training data rather than from G. The basic structure of the GAN model can be visualized in Figure 4. The mathematical minmax optimization (G ∗) of neural networks G and D is as follows:

$$
\begin{aligned}
G^* \in & \arg\min\max V(G, D) \\
= & \arg\min\max \mathbb{E}_{X \sim P_{data(X)}}[\log(D(X))] \\
& + \mathbb{E}_{Z \sim P_{Z(Z)}}[1 - \log(D(G(Z))], \qquad (6)
\end{aligned}
$$

where Z is the input for generator G(Z) with probability distribution PZ and return X with certain probability distribution Pg. The discriminator D(X) estimates the probability that X is from the distribution of training data Pdata. Deep learning has achieved remarkable progress in computer vision and robotics. Moreover, the areas of digital face images and video manipulation are of leading interest because they use the power of GANs, which are capable of producing very realistic results. Nevertheless, without further adjustments such as regularization to encourage greater disentanglement, this technique is difficult to apply in GANs. However, GANs still have challenges in establishing disentangled and controllable syntheses, particularly in the high-resolution domain. Disentangling distinct elements allows us to regulate changes across all factors independently

2.4 PROPOSED ALGORITHM

A. *TRADITIONAL FORENSIC-BASED TECHNIQUES*

- Active techniques require prior knowledge of multimedia for the authentication process.
    - Basically, at the time of multimedia generation, some information is encoded, such as watermarks and digital signatures.
    - Initially, the landmark is extracted from the face images to identify the face region of the person; next, the detected face region is used to extract features.
    - Finally, features are extracted from the detected facial region, and the scores are fused to calculate the final result based on the performance of the classifier according to these features.

B. DEEPFAKES FORENSICS-BASED TECHNIQUES

DeepFake forensics-based approaches are currently a very popular option. This is an active research area. Because of the widespread use of DeepFake technologies, It is relatively simple to manufacture bogus content on the internet. It's quite realistic, and it's impossible to tell it apart from traditional animation. techniques.

- To prevent the harmful effects of DeepFakes, academics have focused their efforts on developing multimedia forensic tools to detect DeepFakes.

- Existing techniques have primarily focused on either spatial and temporal artefacts or data-driven classification. Researchers have recently developed DeepFake detection models using attributes like those shown in Figure 4.
- This section examines these characteristics in order to develop detection algorithms, as well as a list of common methodologies.

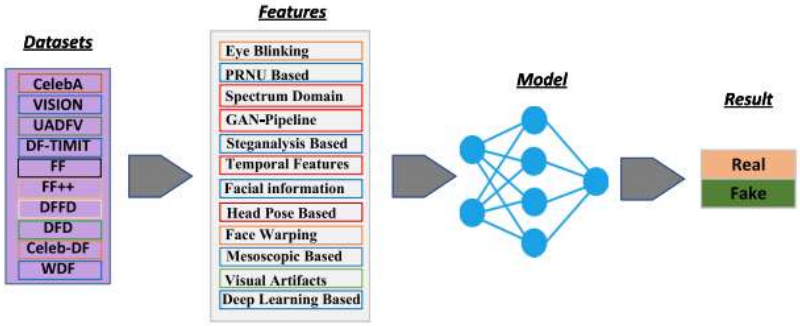

Figure 4. Some important features used for detection.

3. SIMULATION RESULTS

Despite major efforts to improve the visual quality of generated DeepFakes, there are still a few challenges to solve. Generalization, temporal coherence, lighting constraints, lack of realism in eyes and lips, hand movement behaviour, and identity leaking are some of the issues associated with making DeepFakes. The type of dataset provided during training determines the properties of generative models. As a result, after completing training on a certain dataset, the model's output reflects the learned properties (fingerprint). Furthermore, the output quality is influenced by the amount of the dataset used during training. As a result, in order for the model to provide high-quality output, it must be fed a dataset large enough to attain a specific sort of feature. Furthermore, developing a believable model necessitates extensive training. Getting a dataset with relevant material is usually easier; but, finding enough data for a single victim can be tough. It takes time to retrain the model for each new target identification. Figure 5 depicts the fingerprints left by various DeepFake generating models, which can be used to identify them.

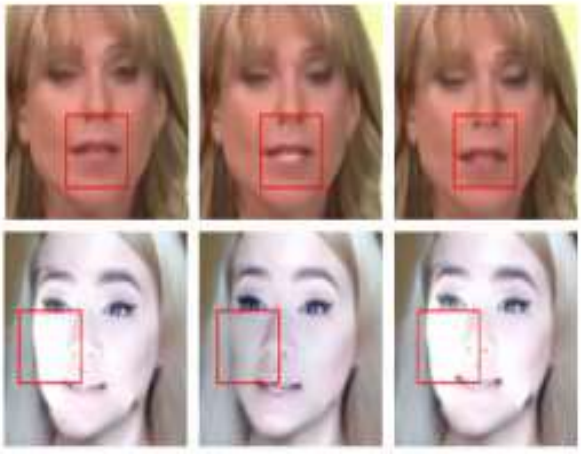

Figure 5. Abnormalities of temporal coherence

Visual anomalies such as flashing and jittering between frames are also problems. Because the DeepFake generation frameworks work on each frame without addressing temporal consistency, several issues occur. Some researchers supply this context to the generator or discriminator, consider temporal coherence losses, employ RNNs, or combine these approaches to address these shortcomings. Figure 5 shows the visible anomalies.

4. CONCLUSION

An innovative and well-liked approach called DeepFake is fully described in this article. The basics, advantages, and drawbacks of DeepFake, GAN-based DeepFake applications are explained. Discussions about DeepFake detection models are also included. The majority of current deep learning-based detection methods fall short in transfer and generalisation, suggesting that multimedia forensics has not yet achieved the peak of its development. Numerous eminent organisations and experts who support the advancement of applied procedures have expressed a great deal of curiosity. The implementation of additional security measures is necessary since maintaining data integrity still requires a significant amount of work. Additionally, experts anticipate a fresh wave of DeepFake propaganda in AI vs. AI conflicts when neither side has an advantage over the other.

5. REFERENCES

[1]. H. Farid, ‘‘Image forgery detection,’’ IEEE Signal Process. Mag., vol. 26, no. 2, pp. 16–25, Mar. 2009
[2]. M. Masood, M. Nawaz, K. M. Malik, A. Javed, and A. Irtaza, ‘‘Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward,’’ 2021, arXiv:2103.00484.
[3]. D. P. Kingma and M. Welling, ‘‘Auto-encoding variational Bayes,’’ in Proc. Int. Conf. Learn. Represent. (ICLR), 2013, pp. 1–14.
[4]. Y. Ma, J. Liu, Y. Liu, H. Fu, Y. Hu, J. Cheng, H. Qi, Y. Wu, J. Zhang, and Y. Zhao, ‘‘Structure and illumination constrained GAN for medical image enhancement,’’ IEEE Trans. Med. Imag., vol. 40, no. 12, pp. 3955–3967, Dec. 2021.
[5]. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, ‘‘Celeb-DF: A large-scale challenging dataset for DeepFake forensics,’’ in Proc. IEEE/CVF C
[6]. D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horvath, E. Bartusiak, J. Yang, D. Guera, F. Zhu, and E. J. Delp, ‘‘Deepfakes detection with automatic face weighting,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2020, pp. 1–9.
[7]. H. T. Le, S. L. Phung, P. B. Chapple, A. Bouzerdoum, C. H. Ritz, and L. C. Tran, ‘‘Deep Gabor neural network for automatic detection of minelike objects in sonar imagery,’’ IEEE Access, vol. 8, pp. 2169–3536, 2020
[8].K. Fukushima, ‘‘Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,’’ Biol. Cybern., vol. 36, no. 4, pp. 193–202, Apr. 1980
[9]. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ‘‘Gradient-based learning applied to document recognition,’’ Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[10]. S. Dong, P. Wang, and K. Abbas, ‘‘A survey on deep learning and its applications,’’ Computer. Science . Rev., vol. 40, May 2021, Art. no. 100379.
[11]. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ‘‘Going deeper with convolutions,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.
[12]. D. Shullani, M. Fontani, M. Iuliani, O. A. Shaya, and A. Piva, ‘‘VISION: A video and image dataset for source identification,’’ EURASIP J. Inf. Secur., vol. 2017, no. 1, pp. 1–16, Dec. 2017.
[13]. Dr.Manjunath R, S.Balaji. Review & analysis of multimedia data miming tasks and models.(2014)
[14]. Dr.Manjunath R, S.Balaji.An Adapative Fuzzy C-means clustering method for retrieval of videos in multi media.