Volume 9, Issue 4, April – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1888

Detecting Fake Images Using Convolutional Neutral Networks - A Deep Learning Approach

1st Servepalli Moushmi Deekshith
Department of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur 522502, Andhra Pradesh, India

2nd Kandepu Niharika
Department of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur 522502, Andhra Pradesh, India

3rd Adapa Akanksha Sri Karthika
Department of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur 522502, Andhra Pradesh, India

4th Gunji Deepika
Department of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur 522502, Andhra Pradesh, India

5th Manoj Wadhwa
Department of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur 522502, Andhra Pradesh, India

Abstract:- One popular generative model with many uses is the Generative Adversarial Network (GAN). Based on this unique concept, recent investigations have shown that it is feasible to produce high-quality fake face photos. The misuse of those fictitious faces in picture manipulation might lead to moral, ethical, and legal issues. To identify fake face images produced by the best method available at the moment, we first propose a Convolutional Neural Network (CNN) based method in this paper [20]. We also present experimental evidence demonstrating that the proposed method can achieve satisfactory results with an average accuracy over 99.4%. To further bolster the logic of our approach, we also offer comparison findings based on a few variations of the suggested CNN design, such as the high pass filter, the quantity of layer groups, and the activation function.

Keywords:- *Fake Image Detection, Deep Learning, Convolutional Neural Networks, Generative Adversarial Networks (GAN).*

I. INTRODUCTION

The ease of altering a picture without noticeable visual artefacts has increased with the swift advancement of image processing technologies. These days, one no longer believes what they see. Over the past 10 years, image forensics have gained a lot of interest, and several forensic techniques based on manually created features have been presented up to this point [4, 15, 16, 19]. Unlike traditional techniques that rely on manually created features, deep learning may use cascaded layers to build hierarchical representations from incoming data in an adaptable manner. Numerous image-related applications, including image style transfer [8, 11], picture super-resolution [13, 18], image inpainting [10, 22], and image steganalysis [6, 21], have used some of the more innovative deep learning models, such CNN and GAN, with remarkable success. For picture forensics, a number of deep learning-based methods have been proposed thus far. For example, Bayar et al. [2] proposed a new CNN architecture to detect several common image manipulations; Rao et al. [17] proposed a CNN-based method to detect image splicing and copy-move; Chen et al. [5] proposed a CNN-based median filtering forensic method; and Choi et al. [7] proposed a CNN-based method to detect composite forgery detection. Recent research has demonstrated that, using a GAN model, it is possible to produce artificial facial pictures with excellent visual quality (see Section 2 for more information). The ability of these false face photos to deceive human eyes makes detecting fake photos a crucial problem for image forensics. Our proposal in this research is to identify the bogus pictures produced by the work using a CNN-based approach [20]. We meticulously plan the CNN architecture in our approach, paying close attention to the activation function, number of layer groups, and high pass filter for the input image. We then present comprehensive experimental results to demonstrate the efficacy and logic of the suggested approach. To the best of our knowledge, no previous research has been done on this forensic issue. This is how the remainder of the paper is structured. Two recent GAN-based face creation works are described in Section 2. The suggested CNN-based detection technique is provided in Section 3. Experiment findings and comments are presented in Section 4. And then, some closing thoughts of this work and the next projects are listed in Section 5.

II. METHODOLOGY AND PROPOSED METHOD

One well-known generative model that generates new samples is generative adversarial networks (GAN) [9], which learns the distribution from high-dimension data. A GAN normally consists of two components: a discriminator and a generator. The maker becomes adept at producing fake data

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1888

that is indistinguishable from the real data, and the discriminator has the ability to discern between false and real input data. During training, they compete with one another until the generator can provide high-quality fake data. A number of GAN-based techniques have been presented recently to produce high-quality fake face photos. For example, Berthelot et al. suggested a unique equilibrium technique in [3] to balance the two components of a GAN to produce aesthetically attractive face pictures. Nevertheless, this technique is limited to creating phoney facial pictures at low resolutions, such 256 × 256. Karras et al. presented a step-by-step method in [20] for building and training GANs to produce high-quality photographs. Fig. 1 shows the progressive method in action. Rather of training the whole GAN on high-resolution photos, they first build a basic GAN training on low-resolution images, and then they progressively add additional layers to the model to make it suitable for high-resolution images throughout the training phase.

The first row of Fig. 8 illustrates how difficult it is to recognise the majority of false face pictures (1024 × 1024) created by this technology with the naked eye, according to the results of the studies. But this approach also yields some less satisfactory findings, as the second row of Fig. 8 shows. In this research, we first provide a way to recognise those high-quality fake face photos produced by the approach [20].

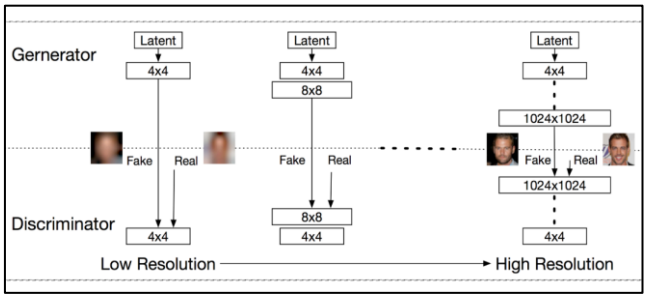

Fig 1 The Progressive Training Strategy Employed in [20]. Here NxN Refers to Layers Operating on Images of NxN Resolution

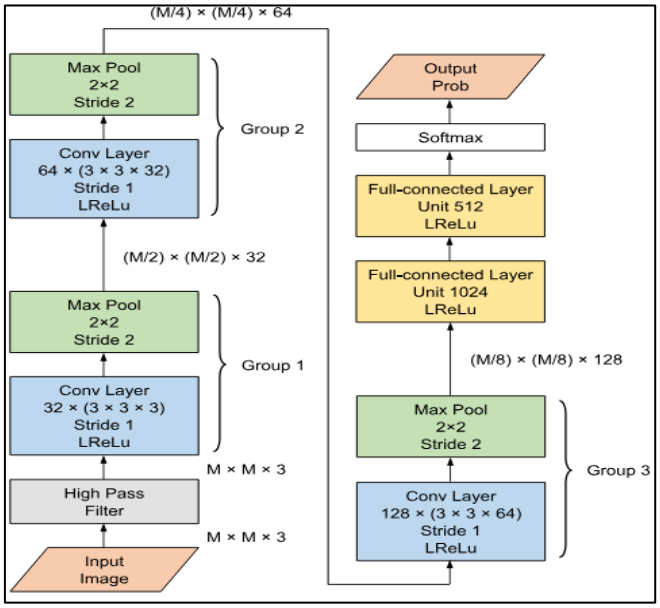

Fig 2 The Proposed Architecture

Given that [20] uses a generator and discriminator that are mostly CNN-based, it makes sense to identify the generated phoney face pictures using a CNN-based technique. To do this, we meticulously create the the suggested CNN model's design, as seen in Fig. 2. An RGB colour picture with a size of M × M × 3 serves as the model input. According to earlier study [14], it is anticipated that the primary distinction between the two types of pictures would be represented on the residual domain since the contents of false and actual facial photos are rather similar. As a result, we first use a high pass filter to convert the input pictures into residuals. Following that, the residuals are routed into three layer groups. A max pooling layer (2 × 2 size, 2 × 2 stride) and a convolutional layer (3 × 3 size, 1 × 1 stride) with LReLu are included in each group. The number of the output feature map of the convolutional layer in the first group is 32, but the equivalent input feature map number is doubled for the subsequent convolutional layers. The last group's output feature maps are then combined and fed into two fully-connected layers. They both have LReLu installed and are made up of 1024 and 512 units, respectively. Lastly, the output probability is generated using the softmax layer.

In our tests, we use Tensorflow [1] to create the suggested CNN model and Adam [12] to train it with a learning rate of 0.0001. Initialising each weight with a truncated Gaussian distribution with a standard deviation of 0.01 and a mean of 0.01. The biases have a zero initialization. In the fully-connected layer, L2 regularisation is enabled with a λ of 0.0005. We train the suggested CNN for 20 epochs using a batch size of 64 during the training phase. We also rotate the training set of data in between epochs.

III. DATASET AND RESULTS

The picture data collection that we used for our research is initially described in this section. Then, we demonstrate the efficacy of the suggested strategy in detecting phoney face photos with a few trials. Furthermore, we carry out comprehensive tests to demonstrate the logic of the suggested model.

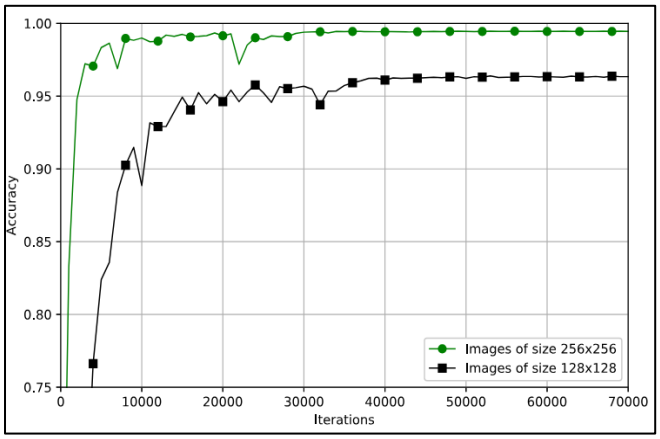

Fig 3 Comparision of Different Image Sizes

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1888

➢ *Dataset:*

We choose 30,000 high-quality fake face pictures from the fake face image database 1 created by [20] and utilise 30,000 real face photos from the CELEBAHQ dataset for our studies. Every image is saved in PNG format and has a resolution of 1024 × 1024. Throughout our trials, Bilinear interpolation is used to resize all pictures to 256 × 256, and lossy JPEG compression with a quality factor of 95 is used to compress them. Lastly, we separate the generated photos into three sets: a validation set consisting of 3,000 pairs of true-fake faces, a test set including 15,000 pairings, and a training set of 12,000 pairs. We divided the training, validation, and test sets three times at random to get findings that were convincing, and we reported the average of those splits in the trials that followed.

➢ *Fake Image Identification*

Finding out if a particular facial image is created or real is our goal in this section. As the blue box in Figure 9 illustrates, we discovered that some background areas in certain phoney face photos appear abnormal. It could improve the effectiveness of detection. We guarantee that each reduced segment primarily comprises some face key-points (such as eyes, nose, and mouse) by cropping a short segment (128 × 128) from each picture in the original image set (256 × 256), as shown in red box in Fig. 9. This reduces the effect of image backgrounds. In conclusion, we have two distinct sets of picture data for our experiments: the original images, which feature the face and backdrop, and the cropped images, which only include the major facial region.

Fig. 3 displays the experimental outcomes that were assessed on the two validation sets. As can be seen in Fig. 3, the suggested model was trained on both the original photos (green line) and After 40,000 iterations, both detection accuracies would be over 95%, and cropped pictures (black line) can converge in around 70,000 iterations. We test the trained model on the test set for a more convincing outcome, and find accuracies of over 99.4% and 96.3% on the original photos and cropped images, respectively. This indicates that even with the background features removed, we can still get good results.

➢ *Comparing the Proposed Model's Variants*

We provide some findings in this part to support the logic of Fig. 2's suggested model. The number of layer groups, the activation function, and the high pass filter are the three components of our model that are taken into account. The subsequent subsections display the related outcomes that were assessed on the validation set.

➢ *High Pass Filter*

The model in this experiment evaluates the three high pass filters listed below.

Fig. 5 presents the same results. As we can see from Fig. 5, out of the three test filters, the suggested model—that is, utilising filter B—can attain the maximum accuracy. Additionally, we note that the model using filter C has the lowest detection accuracy, whereas the model with filter A can obtain comparable results with our suggested model. Additionally, the removal of the high pass filter results in a detection accuracy of about 98%, indicating that the performance of detection may be enhanced by the use of an appropriate high pass filter.

$$\begin{bmatrix} -1 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \quad \begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix}$$

(a) filter A **(b) filter B** **(c) filter C**

Fig 4 Three High Pass Filters

➢ *Number of Layer Groups*

In this experiment, we assess the impact of changing the number of layer groups in the suggested model by one or more. Fig. 6 presents the findings. As can be seen in Fig. 6, the inclusion of one group in the proposed model does not improve detection performance, whilst the removal of one group marginally lowers detection performance. This suggests that the suggested model's use of three layer groups is enough for the problem under investigation.

➢ *Activation Functions*

Another crucial component of CNN is the activation function. The suggested model in this experiment takes six frequently used activation functions into account. These include TanH, ReLu, and four of its variations, such as PReLu, LReLu, ReLu6 and ELU. Figure 7 displays the outcomes of the experiment. We can see from Fig. 7 that PReLu, ELU, and LReLu ReLu may all attain comparable accuracy. Out of the six activation functions, TanH exhibits the weakest performance, and LReLu achieves the highest performance.

IV. CONCLUSION

In this research, we first provide a CNN-based approach to detect phoney face photos created using the most advanced technique [20], and we include comprehensive experimental data to demonstrate the effectiveness of the suggested technique can distinguish between authentic and phoney face photos with good visual quality. Our experimental findings further show that, despite the fact that the existing GAN-based techniques are capable of producing realistic-looking faces (or other visual objects and sceneries), certain evident statistical artefacts will unavoidably be added and may be used as proof of fraudulent ones.

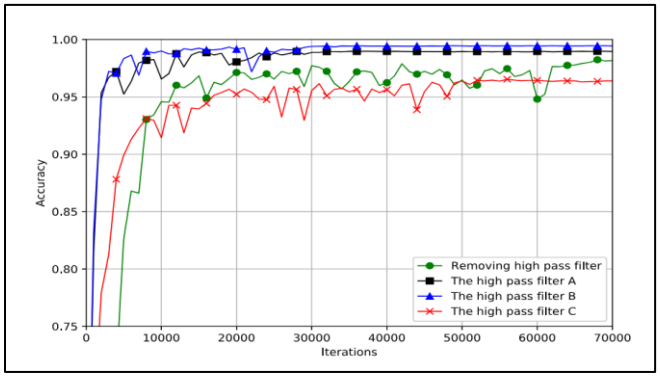

Fig 5 Comparison of Different Image Sizes

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1888

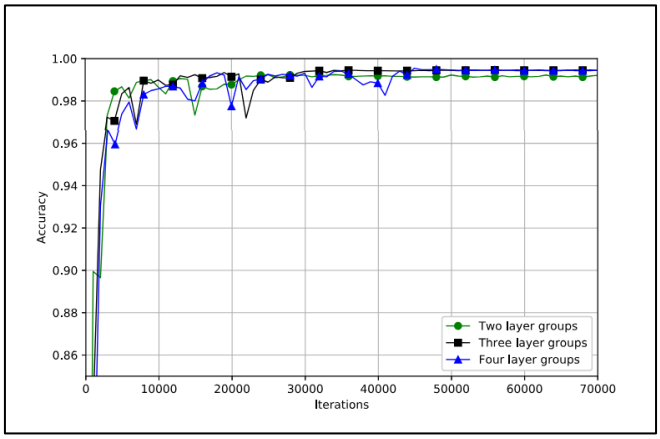

Fig 6 Comparison of Different Number of Groups

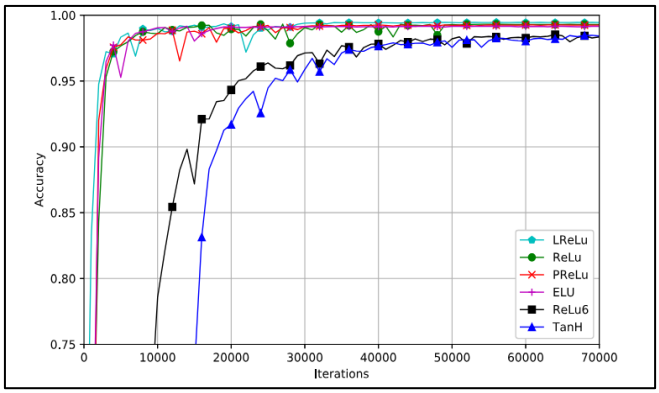

Fig 7 Comparison of Different Activation Functions

We will look at some of the intrinsic artefacts that the GAN in [20] leaves behind in the future for picture forensics. Conversely, we will attempt to suggest a clever face-generation technique that can evade detection.


Fig 8 Fake Face Examples from the Work [20]. The First Row shows examples with a good visual quality, while the Second Shows Ones with a Poor Visual Quality that would be Removed in our Experiments.


Fig 9 Fake Face vs. Background. The Region in the Red Box Includes Some Facial Key-Points; While the Blue Ones are Located at the Background with Poor Visual Artifacts.

ACKNOWLEDGEMENT

This work has been supported by our University research guides and our team members.

REFERENCES

[1]. Elyassami S et al (2022) Fake news detection using ensemble learning and machine learning algorithms. Combating Fake News with Computational Intelligence Techniques. Springer, Cham, pp 149–162
[2]. Ozbay F.A, Alatas B. Fake news detection within online social media using supervised artificial intelligence algorithms.
[3]. Ahmad I., Yousaf M., Youaf S., Ahmad M.O. Fake news detection using machine learning esemble methods.
[4]. AlShariah, N.M., Khader, A., & Saudagar, J. Detecting Fake Images on Social Media using Machine Learning
[5]. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016).
[6]. Belhassen Bayar and Matthew C Stamm. 2016. A deep learning approach to universal image manipulation detection using a new convolutional layer. In Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security. 5–10.
[7]. David Berthelot, Tom Schumm, and Luke Metz. 2017. Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717 (2017).
[8]. Gang Cao, Yao Zhao, Rongrong Ni, and Xuelong Li. 2014. Contrast enhancementbased forensics in digital images. IEEE transactions on information forensics and security 9, 3 (2014), 515–525.
[9]. Jiansheng Chen, Xiangui Kang, Ye Liu, and Z Jane Wang. 2015. Median filtering forensics based on convolutional neural networks. IEEE Signal Processing Letters 22, 11 (2015), 1849–1853.
[10]. Mo Chen, Vahid Sedighi, Mehdi Boroumand, and Jessica Fridrich. 2017. JPEGPhase-Aware Convolutional Neural Network for Steganalysis of JPEG Images. In ACM Workshop on Information Hiding and Multimedia Security. 75–84.
[11]. Hak-Yeol Choi, Han-Ul Jang, Dongkyu Kim, Jeongho Son, Seung-Min Mun, Sunghee Choi, and Heung-Kyu Lee. [n. d.]. Detecting composite image manipulation based on deep neural networks. In IEEE International Conference on Systems, Signals and Image Processing. 1–5.
[12]. Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2017. A learned representation for artistic style. In Proceedings of International Conference on Learning Representations.
[13]. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Advances in neural information processing systems. 2672–2680.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1888

[14]. Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. ACM Transactions on Graphics 36, 4 (2017), 107:1–107:14.
[15]. Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for realtime style transfer and super-resolution. In European Conference on Computer Vision. Springer, 694–711.
[16]. Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. CoRR abs/1412.6980 (2014). arXiv:1412.6980 http://arxiv.org/abs/1412.6980
[17]. Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4681–4690.