IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 3, Ser. 1 (May. – June. 2024), PP 18-24
www.iosrjournals.org

Synth Vision : Robust Attentive Pixel-Wise Gan Generated Fake Face Detection Using Facial Behavior Analysis

Mr. Madhu A,
Assistant Professor, Dept. Of Computer Science And Engineering, Srm Institute Of Science And Technology, Chennai, India

Mamilla Veera Varun,
Dept. Computer Science Engineering Specialization With Big Data Analytics, Srm Institute Of Science And Technology,
Chennai, India

J Sai Mounish,
Dept. Computer Science Engineering Specialization With Big Data Analytics, Srm Institute Of Science And Technology,
Chennai, India

A Dhanush Sai,
Dept. Computer Science Engineering Specialization With Big Data Analytics, Srm Institute Of Science And Technology,
Chennai, India

Abstract-
Image generation has advanced significantly as a result of the generative adversarial networks' (GANs) quick development. Still, the growing realism of generated images as demonstrated by models like StyleGAN and BigGAN, raises concerns regarding national security, social stability, and personal privacy. Through our analysis, we've identified the perceptual degradation issue in existing attack methods when directly applied to these advanced GAN-generated images. Consequently, Considering visual perception, we have created a novel adversarial assault strategy specifically designed for image anti-forensics, especially in the altered color domain. Despite its simplicity, this strategy works incredibly well, fooling forensic detectors based on deep learning as well as those not, with a far greater success rate and significantly better visual quality. Our method uses scene understanding models to extract out-of-context objects from faces. Our method detects anomalous things in facial photographs by utilizing sophisticated object detection and scene understanding techniques. Our research has revealed clear distinctions in contextual cues between actual and synthetic faces. In particular, we found distinct groups of items connected to fake faces as opposed to genuine ones when using scene comprehension and object detection models. Our method surpasses previous approaches, as demonstrated through experiments on fake faces generated by DCGAN. Additionally, we developed an enhanced convolutional neural network (CNN) model tailored for extracting facial features and detecting fake faces. Our experiments indicate promising detection capabilities, particularly in real-world scenarios, with a notable reduction in false positives. DCGAN

Keywords- *Generative Adversarial Networks ,Object Detection, Detection Capabilities, BigGAN, DCGAN, Convolutional Neural Network (CNN)*

Date Of Submission: 24-04-2024 Date Of Acceptance: 04-05-2024

I. Introduction

Neural networks with additional layers have been created as a result of recent developments in deep learning, greatly improving their capacity to learn and generalize from data. But getting precise and trustworthy labeled data to train these models is still an expensive undertaking, especially for computer vision and machine learning applications. To address this challenge, researchers have turned to data augmentation techniques, such as synthesizing additional images to augment the training set, thereby improving the accuracy of image recognition models. The emergence of Generative Adversarial Networks (GANs) has revolutionized the synthesis

of realistic fake faces, posing new challenges for fake image detection methods. As GAN-generated fake faces become increasingly realistic, conventional detection techniques struggle to keep up due to outdated training data. One potential solution is to reduce the reliance on extensive training data by exploring Few-Shot Learning (FSL), which aims to generalize knowledge from limited data. By adapting FSL techniques, fake face detection systems can better cope with evolving generative models, even in scenarios with limited training samples. Despite the proliferation of high-quality image synthesis enabled by GANs, detecting GAN-generated images remains a challenging task in image forensics. Limited research has been conducted in this area, particularly in detecting images manipulated by GAN-based image-to-image translation techniques. Existing studies have shown that while detectors perform well on original images, they may exhibit significant impairments when applied to compressed images from platforms like Twitter. Researchers have proposed two main approaches to detect GAN-generated images: statistical analysis of pixel values and feature extraction, and the identification of discrepancies between fake samples and real data specifications. The Structural Similarity (SSIM) index serves as a valuable metric for assessing perceived image quality, distinguishing between the brightness, contrast, and structural details of an image. Leveraging SSIM assessment can aid in the detection of manipulated images, providing insights into their authenticity and potential alterations.

Information regarding image quality evaluation often revolves around the Structural Similarity Measurement (SSIM), which encompasses three key components: luminance function (l), contrast function (c), and structure comparison function (s). These factors serve as indicators of structural similarity between images. The mean value estimates brightness, standard deviation gauges contrast, and total variation number measures structural resemblance. SSIM operates on a pair of images—an undistorted one and a distorted counterpart—assessing their structural similarity to indicate the quality of the distorted image.In contrast to conventional measurements such as Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR), SSIM more closely resembles human vision when assessing image quality. Geometrically, a vector field of image components which could contain pixels or derived elements like linear coefficients could represent the link between SSIM and conventional metrics. MSE, on the other hand, quantifies the disparity between estimated and original values by squaring pixel differences. It reflects the degree to which estimators deviate from the quantity being estimated.

II. Literature Review

Roose (2022) examines small modular reactors (SMRs), highlighting their smaller size, mobility, and economic advantages compared to traditional nuclear reactors. The paper provides an overview of SMR types, including light water reactors (LWRs), liquid metal-cooled reactors (LMRs), molten salt reactors (MSRs), and gas-cooled reactors (GCRs), emphasizing their structural design, applications, and impacts on power systems.[1] Rombach et al. (2022) present latent diffusion models (LDMs) for high-resolution image synthesis, achieving state-of-the-art results by decomposing the image formation process into denoising autoencoders and utilizing cross-attention layers for conditioning inputs, enabling efficient convolutional synthesis.[2] Pennycook and Rand (2021) examine why individuals believe and share false news online, finding that political beliefs do not solely drive susceptibility; instead, poor truth discernment is associated with lack of careful reasoning and reliance on heuristics, with social media sharing often driven by inattention rather than intentional misinformation dissemination.[3] Singh and Sharma (2022) develop an effective multi-modal method that uses a sentence transformer for text analysis and EfficientNetB0 for visual analysis to identify fake images on microblogging sites. Their model achieves high prediction accuracies on Twitter and Weibo datasets, outperforming other state-of-the-art frameworks.[4]Bird et al. (2023) demonstrate the vulnerability of signature verification systems to false-acceptance attacks by robotic arms and conditional Generative Adversarial Networks (GANs). Their study highlights the need for fine-tuning systems with robotic forgeries to mitigate attack prevalence.[5] Khosravy et al. (2021) analyze the feasibility of model inversion attacks (MIA) on deep learning-based face recognition systems under a gray-box scenario, demonstrating the regeneration of users' face images without access to user information, posing a serious privacy threat.[6] Bonettini et al. (2021) propose utilizing Benford's law divergence values and machine learning techniques for intra-class classification of biometric fingerprint images, achieving high accuracies of 100% with Decision Tree and Convolutional Neural Networks (CNN), and 95.95% and 90.54% with Naïve Bayes and Logistic Regression, respectively.[7] Ramesh et al. (2021)proposed a straightforward method for creating text-to-images in a single shot using a transformer model that autoregressively models text and picture tokens.When examined in a zero-shot way, the technique, despite its simplicity, achieves competitive performance with prior domain-specific models.

[8] Deb et al. (2020) provide an approach that uses facial landmark recognition and superpixel segmentation to inform attack strategies in order to produce adversarial face pictures that are identical to source photos. By showing printouts of created hostile pictures via a camera to trick the recognition model, their method effectively executes adversarial assaults against face recognition systems in the real world.[9] Saharia et al. (2022) introduce Achieving remarkable photorealism and language understanding, introduce Imagen, a text-to-image diffusion model that combines massive transformer language models with diffusion models. With a state-of-the-

art FID score of 7.27 on the COCO dataset and preference from human raters, Imagen beats recent approaches in sample quality and image-text alignment.[10] Chambon et al. (2022) introduce the Medical VDM, which produces high-quality medical images while maintaining key features by utilizing variational diffusion models. Experimental results demonstrate its efficacy with a reconstruction loss of 0.869, diffusion loss of 0.0008, and latent loss of 5.740068×10−5, suggesting potential applications in medical education, diagnosis, and treatment planning.[11]Schneider et al. (2023) explore the application of diffusion models to music generation, presenting a cascading latent diffusion approach capable of generating high-quality stereo music from textual descriptions in real-time on a single consumer GPU. They provide open-source music samples and libraries to support future research in the field.[12]Schneider (2023) explores the potential of diffusion models for audio generation, proposing text-conditional latent audio diffusion models with stacked 1D U-Nets. The work aims for real-time inference on consumer GPUs and offers open-source libraries to facilitate future research in the field.[13] Yi et al. (2021) engineer an AI art model trained on Vincent van Gogh-inspired work, addressing ethical issues in AI-generated artworks. Their model enables style transfer to under-represented individuals by leveraging context from a dataset of nearly 6 billion images, and achieves 98.14% accuracy in distinguishing human vs. AI-created artworks using computer vision models.[14] Guo et al. (2023) introduce ArtVerse, a human-machine collaborative painting paradigm for the metaverse era, leveraging parallel theory to facilitate creation, exploration and evolution. Their framework integrates key technologies to establish decentralized art organizations and demonstrate a new ecology of artistic creation in the metaverse.[15] Sha et al. (2022)create the first comprehensive investigation that examines the accuracy of fake images generated by diffusion models from text to image. Their work proposes universal detection and source attribution strategies by using linguistic and visual modalities to distinguish fraudulent images and connect them to their model source. offering insights into the inherent characteristics of these models and promoting the development of countermeasures.[16] Corvi et al. (2022) propose enhancing AI-generated image recognition through computer vision by generating a synthetic dataset resembling CIFAR-10 with latent diffusion, enabling binary classification between real and AI-generated images using a CNN. Their approach achieves 92.98% accuracy, with interpretability via Gradient Class Activation Mapping revealing the model's focus on small visual imperfections in image backgrounds.[17] Amerini et al. (2019) propose aforensic method that improves performance over single-frame techniques by using optical flow fields to discern between real and false video sequences. Their approach, employing CNN classifiers, shows promising results on the FaceForensics++ dataset.[18]Güera and Delp (2018) investigate the efficacy of pre-trained deepfake detection models on newer datasets, finding the XceptionNet model from Rössler et al. (2019) ineffective, with a 51.31% classification accuracy, particularly struggling to detect deepfake videos (13.16% accuracy), indicating the need for continual testing of detection methods as deepfake technology evolves.[19]Wang et al. (2022) investigate deepfake detection using supervised and self-supervised deep learning models, including transformer-based architectures like DINO and CLIP. Their analysis across multiple datasets reveals that transformer models outperform CNNs, with FaceForensics++ and DFDC datasets demonstrating better generalization capabilities, and image augmentations improving performance.[20]

III. EXISTING SYSTEM

Recent advancements in Artificial Intelligence (AI)-generated images are almost impossible for humans to distinguish from real-life photographs due to the revolutionary nature of synthetic data in image synthesis. In light of the critical need for data authenticity and trustworthiness, this paper recommends applying computer vision techniques to improve human capacity to recognise images created by artificial intelligence.First, a synthetic dataset is constructed using latent diffusion to generate a variety of images that can be compared with actual photos. This dataset is designed to replicate the ten classes of the popular CIFAR-10 dataset. This dataset demonstrates the capability of contemporary picture creation models with its complex visual features, such as realistic reflections in water. The binary classification problem is to differentiate between actual and AI-generated images, which is the essence of the classification task.

Convolutional Neural Networks (CNNs) are used to categorize images into two categories: Real and Fake, in order to accomplish this objective. Following a thorough adjustment of hyperparameters and training of multiple network topologies, the best method yields an astounding 92.98% classification accuracy. Furthermore, this work explores the characteristics essential for classification by applying explainable AI methods with Gradient Class Activation Mapping. Interestingly, the interpretation shows that the key factor in classification is not the actual creatures portrayed, but rather minute visual flaws in the background.To aid in future research, the CIFAKE dataset, which was created specifically for this work, is freely accessible to the scientific community.

.

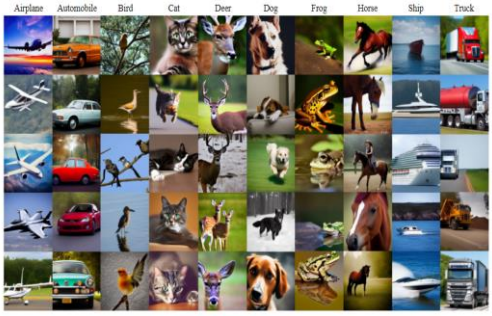


Figure 1: Images from the CIFAR-10 image classification dataset

Class Label	Prompt Modifiers
Airplane	:- aircraft, airplane, fighter, flying, jet, plane
Automobile	:- family, new, sports, vintage
Bird	:- flying, in a tree, indoors, on water, outdoors, walking
Cat	:- indoors, outdoors, walking, running, eating, jumping, sleeping, sitting
Deer	:- herd, in a field, in the forest, outdoors, running, wildlife photography
Dog	:- indoors, outdoors, walking, running, eating, jumping, sleeping, sitting
Frog	:- European, in the forest, on a tree, on the ground, swimming, tropical, wildlife photography
Horse	:- herd, in a field, in the forest, outdoors, running, wildlife photography
Ship	:- at sea, boat, cargo, cruise, on the water, river, sailboat, tug
Truck	:- 18-wheeler, car transport, fire, garbage, heavy goods, lorry, mining, tanker, tow

TABLE 1: Modifiers for Latent Diffusion prompts used to create the synthetic 10-class dataset. "A photograph of {a/an}" appears before each prompt, and moderators are applied equally to all 6000 images.

IV. PROPOSED SYSTEM

The proposed method for fake face detection employs deep learning, treating it as a binary classification task distinguishing between real and fake samples. It introduces a new neural network or loss function for this purpose, drawing inspiration from DCGANs (Deep Convolutional Generative Adversarial Networks). DCGANs enhance standard GANs by utilizing convolutional layers instead of fully connected ones and employing strided convolutions in the discriminator and fractional strided convolutions in the generator. This architecture is particularly effective for spatial correlations in image data, making it a cornerstone for state-of-the-art face generation models. In DCGAN,While the discriminator network learns to distinguish between produced and real images, the generator network creates realistic images from random noise.By iteratively improving each other's performance, DCGAN achieves stable and high-quality image generation. Key features include using convolution layers with stride for upsampling instead of traditional upsampling layers and employing batch normalization to speed up training and enhance stability. For fake face detection, a high-quality face dataset is essential. The method introduces a new loss function to expedite convergence and adopts virtual batch normalization to reduce sample dependence within mini-batches. This approach ensures efficient training and robust performance in detecting fake faces.

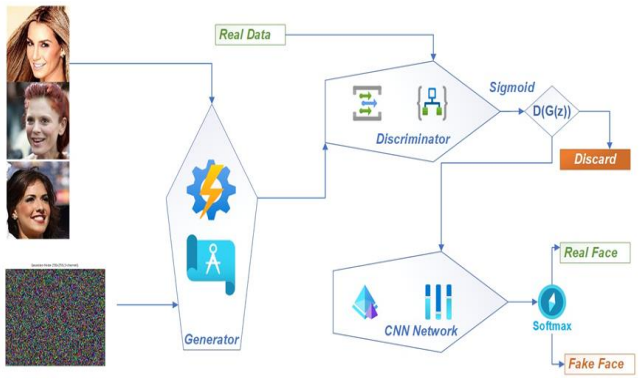


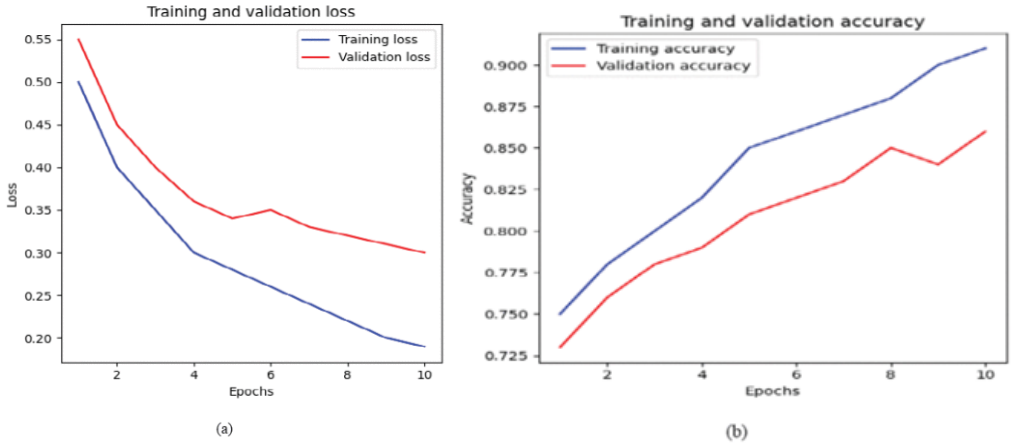
Figure 2: Architecture

The architectural diagram provides a systematic representation of the process for GAN Generated Fake Face Detection using Facial Behavior Analysis here are the modules included :

- **Module 1:** Image Preprocessing Traditional data augmentation encompasses fundamental modifications like horizontal flipping, alterations in color space, and automatic cropping. These techniques address various invariances crucial for image classification models. Recent advancements have expanded augmentation methods to include geometric transformations, color space alterations, kernel flips, image blending, random erasing, increased function spaces, adversarial preparation, shifts in neural architecture, and meta-learning systems. While manual augmentation methods have been conventional, recent experiments focus on leveraging deep neural networks to automatically generate new training samples. Image cropping involves extracting a central patch from an image, a common step in preprocessing for images with combined width and height dimensions. Random cropping, on the other hand, introduces variability into the dataset, impacting interpretation outcomes. Unlike translations, cropping reduces the object's size while preserving the spatial dimension of the image. However, depending on the compression threshold set, this may not always preserve the label information.
- **Module 2:** introduces the DCG-CNN network architecture, which merges improved DCGAN and CNN components to enhance face image expansion and feature extraction. Face images from datasets are fed into the DCGAN to generate diverse facial images with varying shapes and textures, while also generating other images to mitigate false positives. Convolutional layers, pooling layers, and three more convolutional layers make up the CNN structure. From the input photos, these layers gradually extract facial traits.Pooling layers reduce feature dimensionality, optimizing data volume and parameter count without sacrificing feature integrity. A fully connected layer integrates local feature maps for classification, with the final layer containing two neurons indicating the presence of a face image. Network parameters are fine-tuned to achieve a balanced detection rate and minimize false alarms in real-world scenarios.
- **Module 3:** delves into Network Training, where DCGAN utilizes the Adam optimizer. Typically, training commences with the discriminator before training the entire DCGAN. To expedite training, all discriminator layers remain frozen. Introducing multiple discriminators and retraining generators enhances overall network training speed. Ensuring balanced datasets and high-quality face images involves adjusting the proportion of images with various shapes and textures during DCGAN training. Experimentation with different epochs helps identify the most suitable face images. For CNN training, stochastic gradient descent (SGD) serves as the optimizer. Updating the model based on mini-batches instead of individual samples stabilizes convergence. The learning rate, set to 0.01, controls convergence speed. SGD combines gradients and updated weights from previous iterations to update network model weights. contract.

V. RESULT

The proposed system for fake face detection integrates deep learning techniques, particularly inspired by DCGANs (Deep Convolutional Generative Adversarial Networks), to address the binary classification task of distinguishing real from fake facial images. By leveraging convolutional layers and strided convolutions, DCGANs enhance image generation quality and stability. The proposed method introduces a new neural network architecture for fake face detection, incorporating virtual batch normalization to enhance training efficiency and convergence. Additionally, the system employs image preprocessing techniques and a DCG-CNN network architecture for feature extraction and classification. Training involves optimizing the DCGAN with the Adam optimizer and fine-tuning CNN parameters using stochastic gradient descent (SGD).



(a) (b)

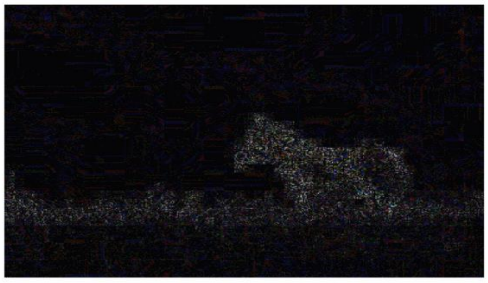
Experimental results using a dataset of 943 images demonstrate a maximum accuracy of 91.83%, indicating the effectiveness of the proposed approach. Graphical representations of accuracy and loss curves validate model performance, with early stopping employed to prevent overfitting. The integration of Error Level Analysis (ELA) grayscale images contributes to improved efficiency and convergence, showcasing promising results in distinguishing between original and modified images. Overall, the proposed system presents a robust framework for fake face detection, offering high accuracy and efficiency in real-world scenarios.

ELA conversion optimizes CNN training efficiency by focusing on non-redundant information in images, highlighting areas with error levels surpassing a set threshold. These distinct features aid in training by emphasizing pixel contrasts, enhancing the model's performance

(a) An Example of an Original Image

Figure 2

After resizing, data is standardized by dividing each RGB value by 255.0, speeding up CNN convergence to the global minimum loss. Labels are then reassigned: '1' for tampered images, '0' for authentic ones. Finally, the dataset is split, with 80% for training and 20% for validation.



(a) ELA Image Results from Figure 2a)

VI. DISCUSSIONS

The proposed system's utilization of deep learning techniques, particularly inspired by DCGANs and CNNs, represents a significant advancement in fake face detection. By leveraging convolutional layers and strided convolutions, the model enhances image generation quality and stability, leading to robust performance in distinguishing between real and fake facial images. The integration of virtual batch normalization further enhances training efficiency and convergence, contributing to the system's effectiveness. Moreover, the inclusion of Error Level Analysis (ELA) grayscale images significantly improves the model's ability to differentiate between original and modified images, ensuring reliable data authentication. Overall, the experimental results demonstrate the system's high accuracy of 91.83%, validating its potential for real-world applications requiring trustworthy image classification.

VII. CONCLUSION

Finally, the project presents an innovative approach to detecting fake faces, drawing inspiration from Generative Adversarial Networks (GANs) and utilizing deep learning techniques. The proposed system achieves a remarkable maximum accuracy of 91.83% on a dataset comprising 943 images, showcasing its effectiveness in discerning between real and fake facial images. By incorporating virtual batch normalization and leveraging Error Level Analysis (ELA) grayscale images, the system demonstrates improved efficiency and convergence. This comprehensive framework offers a robust solution for detecting fake faces with high accuracy, addressing critical concerns surrounding image authenticity and privacy in real-world scenarios.