Deepfake Detection for Image Using hybrid RESNET GAN Framework

Sabal Subedi
Computer Science Department
Idaho State University
Pocatello, ID 83209 USA
sabalsubedi@isu.edu

Abstract

AI is both a blessing and a curse. AI is a blessing because it has expanded a new horizon of learning and working. However, it brought new issues and threads that were misleading and misinformation like a curse. Deepfakes, a technology that is a type of AI engineered to create or modify content to generate convincing yet completely counterfeit images, videos, or audio recordings, is one of the deceptions of AI. With the growing use of social media and social networking, these deceptive visual artifacts generated using some modern image-generative techniques can be a threat to a society and an individual. Thus, deepfake detection has attracted considerable attention in recent years, and numerous novel ideas and methods have been developed to counterfeit it. In this paper, I am presenting a hybrid deepfake detection framework that combines the strengths of Convolution Neural Networks (CNNs) and RESNET 34 within a Generative Adversarial Network (GAN) architecture. Results demonstrate the superior effectiveness of the hybrid model in accurately detecting fake data, representing a significant advancement in facial image recognition and authentication. Evaluation on a benchmark dataset reveals outstanding performance metrics, including a precision of 0.82, recall of 0.78, F1-score of 0.87, and accuracy of 0.89. These findings highlight the hybrid model's robustness. By integrating the generative capabilities of GANs with the discriminative power of ResNet, the proposed model effectively addresses challenges posed by increasingly sophisticated fake face generation techniques. Additionally, the inclusion of Channel-Wise Attention Mechanisms in ResNet34 during the feature extraction phase enhances its performance and overall efficiency.

Keywords: Generative Adversarial Networks (GANs) ; ResNet34 ; Spatial Feature Extraction ; Channel-Wise Attention Mechanism

Introduction

The growing influence of social media and the availability of easy-to-access content with advanced tools and myriad computing infrastructures has made it easier for people to produce deep fakes that can spread misinformation [8]. The creation of propaganda using rapidly advancing technologies can cause panic and chaos. However, like every coin has two faces, deep fakes also have a positive side including the applications in digital aviators, visual effects, creating videos of the last person for their relatives, Snapchat filters, and uploading the episodes without reshooting for movies [7]. Despite that, I focus more on the negative side of the deep fake.

Computer graphics and visualization techniques can generate Deep fakes [3]. This can fuel rapid progress of synthesized image generation and manipulation to a point where it raises significant concerns. At best, this leads to losing trust in digital content but could cause further harm by spreading false information or fake news. Testing the content of digital images/videos would be very crucial as these contents are treated as solid evidence in legal disputes and criminal court cases in standard practice.

Over the past few years, different Machine Learning and Deep Learning areas have been explored to detect deep fakes. Despite substantial research on the subject, there is always potential for improvement in terms of efficiency and efficacy. Also due to the improving and evolving nature of deep fake generation techniques, the datasets are increasingly challenging. Thus, the existing or previous methods may need help to perform effectively [8]. A reliable tool is important to counter the accessible and more sophisticated deep fake technology. Hence, introducing a robust system to detect fake media has become very important in this age of social media [8]. In this paper, the generative capabilities of GANs are combined with the discriminative abilities of ResNet. The model shows its effectiveness in addressing the challenges posed by the increasing fake generation techniques. This paper is organized into six main sections. Section I introduces the research challenge, highlights the significance of the topic, and outlines the objectives of the investigation. Section II provides an overview of the relevant background and related studies. Section III details the methods utilized in the study. Section IV presents the proposed model, including its architecture, design, and implementation. Section V discusses the implementation results and their implications. Finally, Section VI concludes the paper by summarizing the key contributions and suggesting directions for future research.

Research Work

In 1860, the first ever deep fake was developed when a portrait of southern leader John Calhoun was expertly al-

tered for propaganda by swapping his head out for the US President. The typical manipulations are done by splicing, painting, and copy-moving the items inside or between two photos. Then post-processing such as scaling, rotating, and color modification are performed to enhance the visual appeal [9]. In addition to Convolutional methods of manipulation, a range of automated procedures, are now available due to the development of computer graphics and ML/DL techniques, which have improved the semantic consistency.

Over the years, though deepfake detection has become a difficult problem to tackle, many researchers have proposed numerous models. In [6], the authors presented a deep convolution GAN detection model as a solution to the challenge, of distinguishing synthetic data with a very realistic appearance which is so hard to spot with the naked human eye. The study presented methods used to implement deepfakes and included the main deepfake manipulation and detection techniques. In [2], the paper introduced a pioneering hybrid deep learning model, which merges the capabilities of GANs and the Residual Neural Network (RESNET) architecture to detect fake faces. The authors also portrayed the comparative analysis against established pre-trained models such as VGG16 and RESTNET 50.

The authors in [1] proposed a novel proactive DeepFake detection technique that leverages the GAN-based visible watermarking. They proposed a constructive regularization added to the GAN's loss function that embeds a unique watermark to the assigned location of the generated fake image. Then, these watermarks are detected by the SOTA DeepFake detector. Likewise in [4], the study presented a novel class of simple counterattacks that removes the indicative artifacts, the GAN fingerprint, directly from the frequency spectrum of a generated image. The study explored different realizations of this removal, ranging from high-frequency filtering to more nuanced frequency-peak cleansing. The study showed that an adversary can often remove GAN fingerprints and thus evade the detection of generated images. The author in [10] examines various deep learning approaches to detect AI-generated fake faces, emphasizing the necessity of robust detection systems to keep pace with rapidly evolving multimedia manipulation technologies. The review explores methods such as Convolutional Neural Networks (CNNs), Xception Networks, Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) models, aiming to develop more accurate and efficient systems for identifying deepfakes. These strategies leverage spatial and temporal features to improve detection precision and reliability.

In [10], the authors introduced G-JOB GAN, a machine learning model based on GAN, which generates highly realistic images and achieves a 95.7% accuracy in detecting realistic generated images. The same model can also detect fake images with similar probability. The study also presented a comparative analysis against several GAN architectures, including Style GAN, Pro GAN, and the original GAN, and claims that the model outperformed all others

Approach

Residual Neural Network

Residual Neural Network (ResNet) is a deep learning architecture introduced in [5], widely recognized for its exceptional performance in computer vision tasks and its success in achieving state-of-the-art results across various image recognition challenges. The key innovation in ResNet is the incorporation of residual connections], which facilitate the efficient training of very deep neural networks.

The ResNet architecture is built around a series of convolutional layers and residual blocks. A residual block comprises multiple convolutional layers with shortcut connections that bypass these layers [2]. This design allows the network to learn residual functions representing the difference between the input and the target output, thereby streamlining the learning process.

As illustrated in Figure 1, the ResNet architecture includes multiple components such as convolutional layers, residual blocks, and an output layer. The input, typically an initial image or feature map, is processed through a convolutional layer that applies learnable filters to extract features. Each residual block contains two or more convolutional layers, with shortcut connections enabling the input to bypass these layers. The output of the convolutional layers is combined with the original input via these connections, allowing the network to learn the residual function—essentially, the difference between the input and the target output. This approach significantly enhances the training efficiency of very deep networks.

Channel-Wise Attention Channel-Wise Attention Mechanism represents a significant advancement in deep learning architectures, particularly in convolutional neural networks (CNNs), where identifying intricate patterns is essential. These mechanisms enhance the feature extraction process by selectively emphasizing relevant feature channels while minimizing the influence of noise and irrelevant data.

Integrating Channel-Wise Attention Mechanisms into ResNet32 [11], a widely used CNN architecture known for its deep layers and skip connections, brings several performance benefits. By enabling the network to focus on critical features, these mechanisms ensure the accurate identification of key attributes necessary for distinguishing between genuine and altered images, such as in the task of fake face detection. Furthermore, they allow for adaptive feature representation, enabling the model to adjust dynamically to input variations, thereby improving its ability to capture subtle differences. This adaptability enhances the model's discriminative capabilities and strengthens its robustness against adversarial perturbations and unseen data which significantly boosts ResNet34's effectiveness in tasks requiring precise feature extraction, such as detecting fake faces. Typically, after several residual blocks, the network employs global average pooling, which condenses spatial dimensions by calculating the average value of each feature map [11]. This operation aggregates information across the entire image, producing a compact feature representation. These pooled features are then processed by a fully connected layer or a softmax layer to generate the final output, such as class

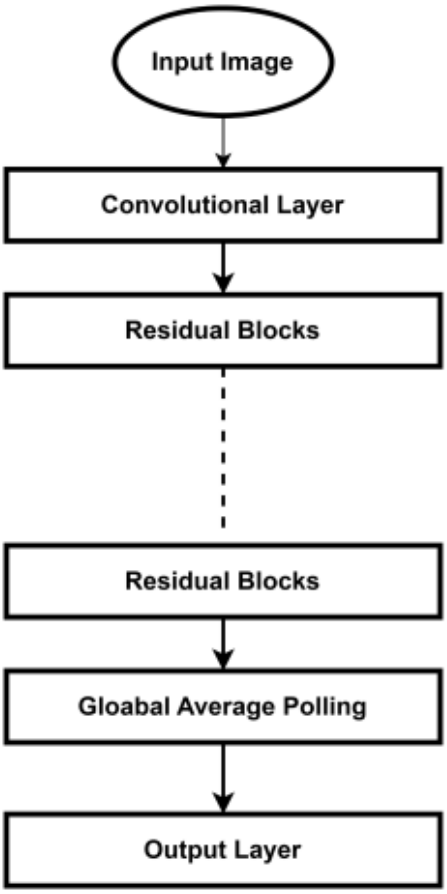

Figure 1: Architecture of RESNET

probabilities. This combination of attention mechanisms and pooling strategies ensures efficient and accurate feature utilization for diverse applications.

GAN

GAN is a type of deep learning architecture that is used for generating new data samples such as images generated in [2]. A typical GAN consists of two components: a generator and a discriminator, where both networks compete with each other as shown in 2. The generator attempts to generate fake data that looks real by learning the features from the real data and acts as the heart of GAN. The discriminator evaluates the generated data with the real data and classifies whether the generated data looks real or not. Then the discriminator provides the feedback to the generator to improve its data generation. The main goal of the generator is to trick the discriminator.

Proposed Model and Dataset

This section provides a detailed explanation of the proposed model and its approach to identifying real and fake faces. The model addresses a critical challenge in face recognition, offering a reliable method for distinguishing between

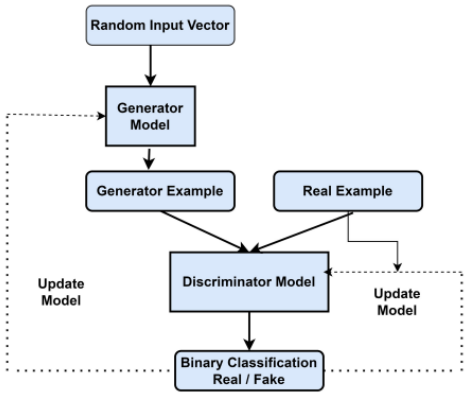

Figure 2: Architecture of GAN

Figure 3: Sample dataset

authentic and fraudulent images. Its application could be valuable in fields such as security and criminal investigations due to its effectiveness in detecting fake faces. Key aspects of the model, including the integration of machine learning, deep learning techniques, and pre-trained models, will be summarized. Additionally, the methodologies used in the study will be reviewed, highlighting both their benefits and potential limitations.

Dataset

The dataset used in this study was found in the Kaggle named ”real and fake faces” and contains 66,497 real face images as shown in 3 and 40,000 fake images. This dataset is frequently utilized to evaluate the performance of various models designed to differentiate between real and fake faces.

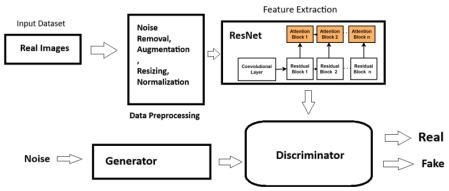

Figure 4: Architecture of proposed model

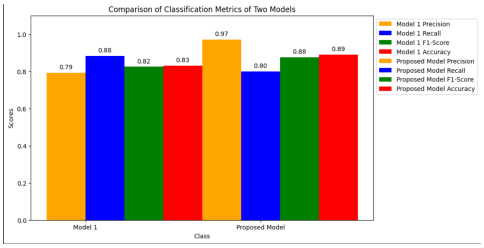

Figure 5: Comparision of classification metrics of compared vs proposed model

Proposed Model and Architecture

The proposed architecture is built around the Generative Adversarial Network (GAN) and consists of 5 steps. The architecture of the proposed model is shown in 4. Data preprocessing is the initial step. First, data cleaning is applied to clean the data from any corrupt or mislabeled images. Any problematic images are removed to ensure data integrity. Then. data augmentation techniques are used to increase the diversity of data and data size such as scaling, crop, flipping, normalization, and rotation. This help to increase model robustness and effective against unseen data. Now, the DataLoader, a crucial utility that helps with loading data efficiently in a way that is compatible with PyTorch's training and evaluation pipelines, is used to create a training data pipeline. Batch loading, shuffling, and parallel data loading are achieved using DataLoader. The training data in pipeline, now, has images of size 64 X 64 pixels, and these pixels are scaled in the range of [0, 1] for faster convergence during the training and to prevent issues related to different pixel value scales. In second, the standard ResNet34 model, a pre-trained Convolutional Neural Network (CNN) was modified to optimize its suitability for a specific analytical task. A ResNet34 model serves as foundational CNN architecture throughout the feature extraction stage. Attention mechanisms are incorporated to capture attention-weighted feature representations. Additional convolution layers were introduced to enhance the both training and predictive efficiency. Furthermore, the kernel sizes were optimized to capture nuanced patterns in the data. Dropout, regularization, and batch normalization were deployed to mitigate the overfitting risks and enhance the model's generalization ability.

Third, the random noise of latent size was created to generate fake faces. These generated faces together with the real faces are fed as input to the discriminator. The discriminator network tries to distinguish between real and fake images. The pre-trained model, ResNet34 is integrated at the GAN's discriminator. The features extracted by the pre-trained model are crucial for discriminator to make the decision. The fourth step is training. The hybrid model is trained using a custom loss function that extends PyTorch's (nn.Moddule) and implements a weighted Binary Cross-Entropy (BCE) loss.

Finally, the fifth phase evaluates the hybrid model's performance using a separate validation or test dataset. Metrics like accuracy, precision, recall, and F1-score are computed to measure the model's ability to effectively distinguish between real and fake faces. This comprehensive evaluation ensures the model's robustness and reliability in detecting fake faces

Result and Analysis

In this section, the results obtained from the proposed model are presented. In this section, the obtained results of many experiments and the proposed model are introduced, but initially, the section briefly describes the different measures used to evaluate the performance of these models.

Sensitivity (Recall): sensitivity measures the proportion of true positives that are correctly identified as such. In other words, it is the probability that a test will correctly identify a positive case.

Sensitivity = TP/(TP + FN)

where TP is True Positive, FN is False Negative. Precision: Precision measures the fraction of positive predictions that are actually positive

Precision = TP/TP + FP.

where FP is False Positive. Accuracy: accuracy measures the fraction of predictions that are correct, regardless of whether they are positive or negative

Accuracy = (TP + TN)/(TP + FP + FN + TN)

F1 Measure: is a weighted average of precision and recall. It is calculated by taking the harmonic mean of precision and recall.

F1Measure = 2 x (precision x Recall)/(precision + Recall)

In order to compare the results of the proposed model, a relevant paper and its results in [2] are considered. The model presented in [2] uses a hybrid of a similar kind but ResNet50 as a pre-trained model and the classification metrics are present in 5. Table 1shows the results yielded from [2], ResNet50 is used to classify the real faces and fake faces. The results of the proposed model is presented in 1. The hybrid model underwent training over 50 epochs, each spanning approximately 2 hours on a Google colab equipped with .12.7 GB of RAM. Comparing these result, we can see that the proposed model outperformed the model in [2] on all other metrics except the Recall. Typically, precision and recall, together, give insight into the trade-off between correctly identifying positive cases and avoiding false positives. Thus, improving recall might decrease precision and vice-versa. Figure 6 shows the generator loss and discriminator loss over the epochs. Both the generator and discriminator are improving over time, but there are signs of volatility

Table 1: Benchmark on compared vs proposed model

	Accuracy	Precision	Recall	F1
Compared Model	0.8298	0.7916	0.8824	0.8345
Proposed Model	0.8906	0.9709	0.8000	0.8767

.

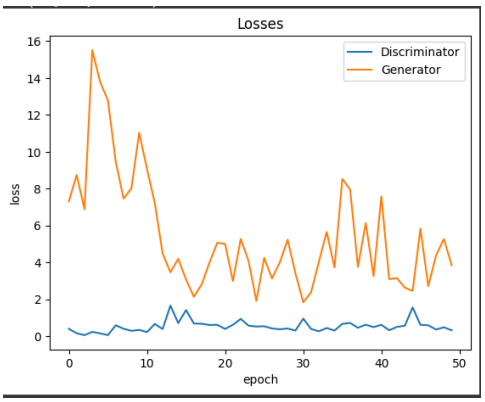

Figure 6: Generator loss vs discriminator loss

.

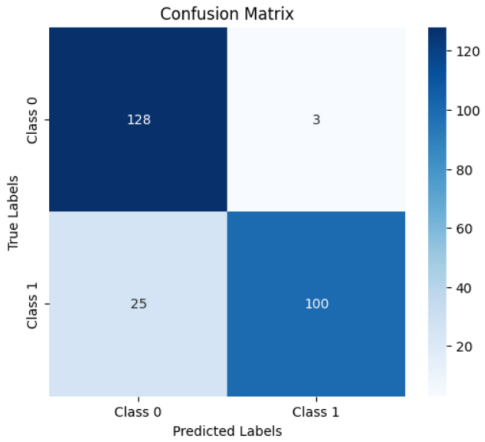

Figure 7: Confusion matrix

around epoch 40. This behavior could indicate challenges in maintaining a balanced training dynamic between the two networks, possibly requiring some fine-tuning.

The model has high precision which means that it predicts a positive class almost correctly as shown in Figure [confusion]. However, the recall could be improved i.e. it misses 20% of the actual positives. The model's accuracy and F1-score shows a solid balance between the precision and recall.Depending on the application, it may be worth tuning the model to improve the recall.

In many cases, a hybrid model, like a combination of GAN and ResNet34, yield better results than a single architecture like RESNET34. The qualities of both components working together give rise to this advantage. The ability of GAN to generate synthetic data while mimicking the dataset combined with the deep design of ResNet34 that excels in extracting the detailed features provides more diversified and informative features for classification problems. Furthermore, data augmentation fills the gap in insufficient training data and diversity which increases the model's efficiency to classify the real and face data. The model shows its effectiveness in addressing the increasing fake generation techniques and its challenges. Thus, the commitment to continuous improvement to enhance the model's effectiveness and reliability for practical cases like social content moderation, identify verification, and cybersecurity is going on.

Conclusion

In this research, a novel hybrid model is proposed that leverages deep learning to address the growing challenge of synthetic data, such as altered or fabricated images. The study integrates the powerful feature extraction capabilities of the ResNet architecture, enhanced by a Channel-Wise Attention Mechanism, and the generative capabilities of GANs to create a robust and accurate classifier. This model holds significant potential to address critical domains such as social media content moderation, digital identity verification, and cybersecurity. Its ability to determine the authenticity of visual content is a crucial advancement in combating the proliferation of deepfakes and manipulated media.

The proposed model's architecture has been carefully designed to balance computational efficiency with high performance, making it suitable for deployment in real-world applications. Extensive experiments demonstrate its effectiveness across diverse datasets, showcasing its adaptability and generalization capabilities.

Future work on this research can include exploring cross-database evaluation to test the model's robustness against unseen datasets, optimizing its performance through fine-tuning with diverse and larger datasets, and investigating new techniques to enhance evaluation metrics further. Additionally, extending the model to handle real-time detection, improving its interpretability for better decision-making, and exploring lightweight versions for deployment on edge devices could provide valuable directions for future enhancements.

References

[1] Ajita Rattani Aakash Varma Nadimpalli, *Proactive deepfake detection using gan-based visible watermarking*, ACM Journals (2024). Accessed: 2024-09-15.

[2] Ajita RattaniAuthors Info Claims Aakash Varma Nadimpalli, *Proactive deepfake detection using gan-based visible watermarking*, IEEE Explorer (2023). Accessed: 2024-09-12.

[3] Luisa Verdoliva Christian Riess Justus Thies Matthias Nießner1 Andreas Rossler Davide Cozzolino, *Faceforensics++: Learning to detect manipulated facial images*, Scientific Reports (2019). Accessed: 2024-10-12.

[4] R. Chauhan, *Deep learning-based methods for detecting generated fake faces*, Authorea Preprints (2023). Accessed: 2024-09-28.

[5] R. Vera-Rodriguez V. Lopes H. Proença J. C. Neves R. Tolosana and J. Fierrez, *Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection*, IEEE Explorer (2019). Accessed: 2024-09-28.

[6] Hitesh Kumar Sharma Prerti Manoj Kumar, *A gan-based model of deepfake detection in social media*, ScienceDirect (2023). Accessed: 2024-09-18.

[7] Tareq Al-shami Tareq Al-shami R Saravana Ram M Vinoth Kumar, *Deep fake detection using computer vision-based deep neural network with pairwise learning*, ResearchGate (2022). Accessed: 2024-09-17.

[8] Rashid Amin Jaroslav Frnda-Aida Mustapha Asma Hassan Alshehri Rimsha Rafique Rahma Gantassi, *Deep fake detection and classification using error-level analysis and deep learning*, Scientific Reports **7422** (2023). Accessed: 2024-11-11.

[9] Rashid Amin Samina Khalid Sultan S. Alshamrani Abdullah Alshehri Shumaila Aleem Noor ul Huda, *Machine learning algorithms for depression: Diagnosis, insights, and research directions*, MDPI (2022). Accessed: 2024-09-10.

[10] Galamo Monkam; Weifeng Xu; Jie Yan, *A gan-based approach to detect ai-generated images*, IEEE Explorer (2023). Accessed: 2024-09-8.

[11] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola, *Dive into deep learning*, Self-published, online resource, 2021. Accessed: 2024-12-12.