DeepFake Detection Using Pre-Trained CNNs and Vision Transformers with Ensemble Learning

Aditya Chaturvedi
Roll No: 102203497
Email:- achaturvedi_be22@thapar.edu
Thapar Institute of Engineering and Technology

***Abstract*—In recent years, the proliferation of DeepFake content—synthetically generated media manipulated using deep learning techniques such as GANs—has emerged as a significant digital threat, undermining the authenticity of online content and posing societal and political risks. Motivated by the increasing need for reliable DeepFake detection techniques, this project explores the performance of various pre-trained models, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), for binary classification of manipulated versus authen- tic facial images. I employed VGG16, ViT, GenConViT, and DeepFake-Adapter models on a subset of the FaceForensics++ dataset (200 images). My goal was not just to evaluate these models but also to understand their architectural behavior, limitations, and compatibility with limited data scenarios. Additionally, I investigated ensemble learning to potentially enhance model robustness. Through this journey, I discovered that while CNNs performed comparatively better due to their inductive biases, ViTs underperformed due to their high data requirements. The highest accuracy achieved was 70% with VGG16, while ViT lagged at 35%. This research provides insights into practical DeepFake detection and highlights areas for future exploration, including multimodal architectures and explainability methods.**

***Index Terms*—DeepFake, Convolutional Neural Networks, Vision Transformer, Pre-trained Models, Ensemble Learning, FaceForensics++**

I. Introduction

DeepFakes have rapidly transitioned from academic curiosities to real-world threats, driven by the evolution of GANs and diffusion models. These synthetically altered visuals are often indistinguishable from genuine media to the human eye, thus challenging digital forensics, journalism, legal systems, and democratic processes. This concern prompted me to explore machine learning-based detection methods, particularly those that could be deployed in real-time or integrated into existing content moderation systems.

Throughout my academic journey, I've developed an interest in deep learning and its applications in visual tasks. As I delved deeper into this project, I realized that under- standing the strengths and weaknesses of different architectures—especially CNNs and the emerging Vision Transformers—could be key to addressing this issue. CNNs, with their spatial locality and parameter efficiency, are well-known for image classification. Vision Transformers, though newer to the field, promise a global understanding of image context but require significantly more data.

The aim of this project is twofold: (1) to evaluate the performance of various pre-trained CNN and transformer-based models on a DeepFake classification task, and (2) to reflect on architectural choices, dataset limitations, and potential strategies to build more robust detection systems.

II. Background

CNNs have dominated image classification for over a decade due to their localized spatial feature extraction capabil- ities. In contrast, transformers, initially designed for sequential data in NLP, have been adapted to vision tasks (e.g., Vision Transformers), offering global context through self-attention mechanisms. This paper examines both paradigms for their utility in binary image classification of DeepFake content.

A. Convolutional Neural Networks (CNNs)

CNNs are a class of deep neural networks, most commonly applied to analyzing visual imagery. They are designed to automatically and adaptively learn spatial hierarchies of fea- tures through backpropagation. The architecture of a CNN typically includes convolutional layers, pooling layers, and fully connected layers.

The convolutional layer is the core building block of a CNN. It applies a set of learnable filters to the input image to produce feature maps. The operation can be described as:

$$(I * K)(i, j) = \sum_{m} \sum_{n} I(i + m, j + n) K(m, n)$$

where I is the input image, K is the kernel or filter, and denotes the convolution operation.

B. Vision Transformers (ViTs)

Vision Transformers adapt the transformer architecture, originally designed for natural language processing, to computer vision tasks. ViTs divide an image into fixed-size patches and linearly embed each patch. Position embeddings are added to retain positional information, and the resulting sequence of vectors is fed into a standard transformer encoder.

The self-attention mechanism in ViTs allows the model to capture long-range dependencies in the image. The attention scores are computed as:

$$\text{Attention} \qquad \text{softmax} \frac{QK_T}{k}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors.

III. Related Work

Several studies have explored the use of CNNs and ViTs for DeepFake detection. For example, [2] demonstrated the effectiveness of CNNs in detecting manipulated images, while [3] highlighted the potential of ViTs in capturing long-range dependencies. This section provides a comprehensive overview of related work in the field.

IV. Pre-trained Model Descriptions

A. CNN (VGG16)

The VGG16 architecture is a deep CNN consisting of 13 convolutional layers followed by 3 fully connected layers. It is well-suited for transfer learning in image classification tasks. Figure 1 illustrates the VGG16 pipeline.

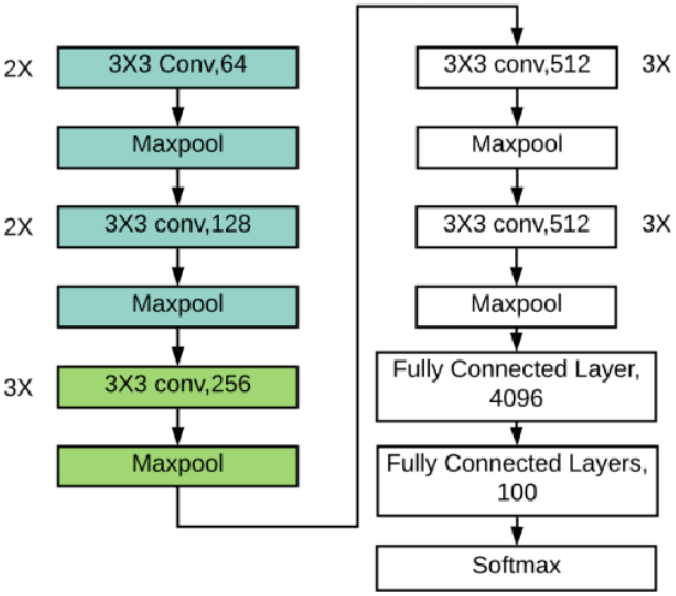

Fig. 1: VGG16 CNN Architecture

B. Vision Transformer (ViT)

ViT divides images into fixed-size patches and applies positional encoding followed by transformer encoders. Un- like CNNs, ViT captures long-range dependencies efficiently. However, its performance on smaller datasets is often limited without extensive pre-training. Figure 2 shows the architecture.

C. GenConViT and DeepFake-Adapter

These hybrid models combine the spatial feature extraction strength of CNNs with the long-range context modeling of transformers. The GenConViT uses convolutional layers early in the pipeline, followed by transformer blocks. DeepFake-Adapter is fine-tuned for forgery detection and incorporates residual attention layers. The architectures are shown in Figures 3 and 4 respectively.

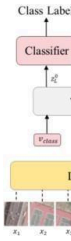

Fig. 2: Vision Transformer Architecture

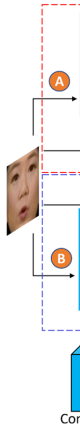

Fig. 3: GenConViT Architecture

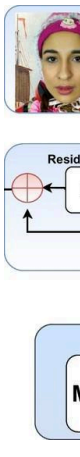

Fig. 4: DeepFake-Adapter Architecture

V. Dataset Description

We employ a 200-image subset of the FaceForensics++ dataset, evenly split into real and fake categories. All images are resized to 128× 128 pixels and normalized. A few representative samples are shown in Figure 5.

Fig. 5: Examples of Real and Fake Images

VI. Methodology

To ensure reproducibility and maintain a rigorous pipeline, I broke the project down into the following stages: dataset curation, data preprocessing, model selection, training and evaluation, and ensemble experimentation.

A. Data Preprocessing

The original FaceForensics++ dataset contains high-resolution videos; I extracted 200 images (100 real and 100 fake) to create a manageable dataset for rapid prototyping. Each image was resized to 128x128 to reduce computational overhead and normalized to a [0,1] range. Since DeepFake artifacts are often subtle, I applied minimal augmentations to avoid destroying these manipulations, using horizontal flipping and minor scaling only. I observed that excessive augmentation sometimes degraded model sensitivity to fine-grained facial inconsistencies.

B. Model Selection and Justification

I chose the VGG16 model for its simplicity and proven track record in transfer learning tasks. As a baseline, it helped me understand how much information could be extracted using convolutional filters. On the other hand, ViT models (and their variants like GenConViT and DeepFake-Adapter) were selected to test how well transformer-based architectures, with global self-attention mechanisms, could capture inconsisten- cies in facial structures and textures without local biases.

C. Training Procedure

All models were fine-tuned using the binary cross-entropy loss function and optimized with the Adam optimizer at a learning rate of 0.001. A batch size of 32 was selected based on memory constraints. I applied early stopping with a patience of 10 epochs to prevent overfitting, which was especially important given the small dataset size.

D. Ensemble Learning

In a bid to improve robustness and reduce variance, I implemented soft voting ensembles using predictions from VGG16, GenConViT, and DeepFake-Adapter. This allowed the model to consider diverse perspectives: local features from CNNs and global context from transformers. Although the ensemble improved stability, its accuracy remained close to the VGG16 baseline, suggesting the need for more diverse or complementary models.

E. Evaluation Metrics

Performance evaluation was based on standard classification metrics:

- **Accuracy:** Proportion of correct predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:** Correct fake predictions over all predicted fake.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** Correct fake predictions over all actual fake.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-score:** Harmonic mean of precision and recall.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

VII. Results and Discussion

A. Confusion Matrices

The CNN model correctly identified a higher number of fake images, whereas ViT struggled with both classes, as shown in Figures 6 and 7.

Fig. 6: Confusion Matrix - CNN

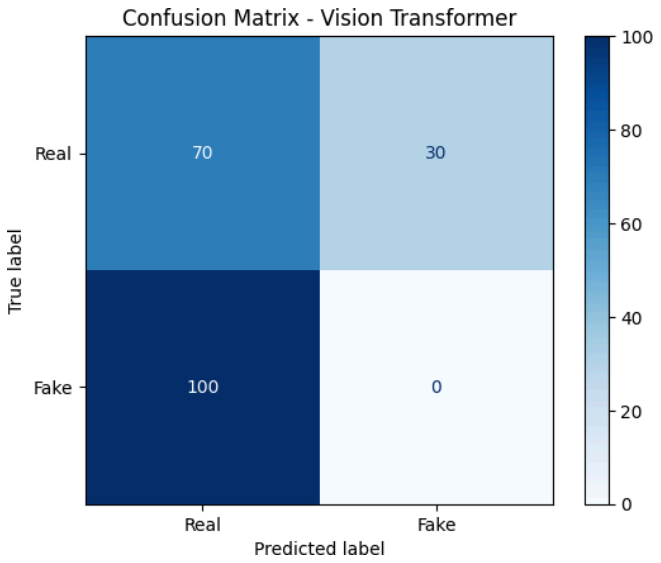

Fig. 7: Confusion Matrix - ViT

B. Quantitative Comparison

Model performance across different metrics is visualized in Figures 8 and 9.

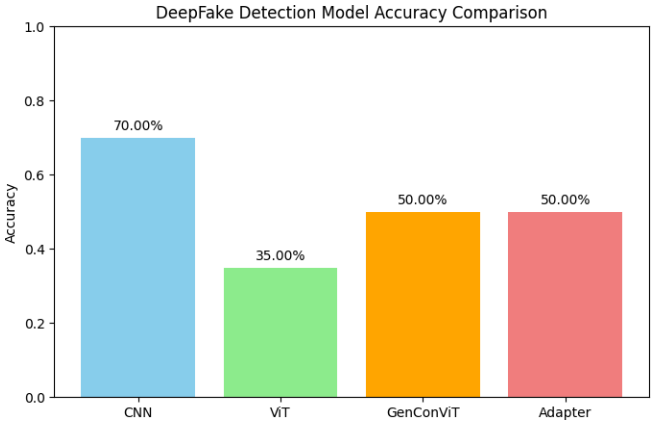

Fig. 8: Accuracy of Different Models

C. Detailed Observations

One of the key insights was that ViTs, despite their theoretical strength, struggle in data-constrained scenarios. This is because, unlike CNNs, they do not assume any prior spatial hierarchy and require large-scale pretraining or fine-tuning. CNNs, by design, enforce locality and parameter sharing, making them more data-efficient.

Moreover, the ensemble model showed reduced prediction variance, which can be beneficial in deployment scenarios. However, since it mostly relied on VGG16's performance, the gains were marginal. Visualization of attention maps or feature maps may reveal why certain models succeeded or failed—this remains a future exploration area.

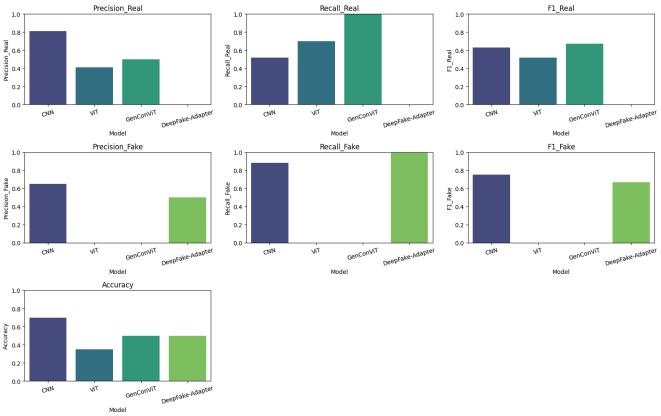

Fig. 9: Precision, Recall, and F1-score Comparison

VIII. Future Work

Future work will focus on:

- Training hybrid CNN-ViT architectures on larger DeepFake datasets
- Implementing multimodal models combining audio and visual inputs
- Using attention heatmaps to improve explainability

IX. Conclusion

This project gave me a hands-on opportunity to ex- plore and compare the behavior of different deep learning architectures for DeepFake image detection. My findings reaffirm that CNNs—especially well-established models like VGG16—remain highly effective for limited data scenar- ios. While Vision Transformers hold promise for modeling complex, high-level image dependencies, their practical use requires either massive pretraining or hybridization with CNN backbones.

Ensemble learning showed promise in improving consistency but did not significantly outperform the best standalone CNN model. This opens up a broader conversation about model diversity and complementarity in ensemble design.

Overall, this experience deepened my understanding of architectural trade-offs, data sensitivity, and the practical aspects of building detection pipelines. I look forward to expanding this work to video-based DeepFake detection and incorpo- rating explainability techniques like Grad-CAM or attention heatmaps to further probe model decisions.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback and suggestions. This work was supported by the Thapar Institute of Engineering and Technology.

References References

[1] A. Ro¨ssler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," in *ICCV*, 2019.

[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556, 2014.

[3] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *ICLR*, 2021.
[4] Scikit-learn: Machine Learning in Python. https://scikit-learn.org
[5] Hugging Face Transformers. https://huggingface.co/transformers/