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

Ashutosh Gupta
Roll No: 102203230
Email: guptaashutosh8950@gmail.com
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 authentic facial images. I employed VGG16, ViT, GenConViT, and DeepFake-Adapter models on a subset of the FaceForensics++ dataset (200 images). The goal was to evaluate these models and understand their architectural behavior, limitations, and compatibility with limited data scenarios. Additionally, ensemble learning was investigated to potentially enhance model robustness. 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 Transformers, 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 project investigates machine learning-based detection methods suitable for integration into real-time or moderation pipelines.

II. Background

A. Convolutional Neural Networks (CNNs)

CNNs are designed for visual tasks using convolutional, pooling, and fully connected layers. The convolution operation is represented as:

$$(I * K)(i, j) = \sum_m \sum_n I(i+m, j+n)K(m, n) \qquad (1)$$

where I is the input image and K is the kernel.

B. Vision Transformers (ViTs)

ViTs divide images into fixed-size patches and feed them into transformer encoders with self-attention mechanisms. The attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2)$$

III. Related Work

Previous studies such as [1] and [3] show CNNs' effectiveness in DeepFake detection and the theoretical potential of ViTs. However, performance depends heavily on dataset size and architecture tuning.

IV. Pre-Trained Models

A. VGG16

VGG16 is a deep CNN with 13 convolutional and 3 fully connected layers, well-suited for transfer learning.

B. Vision Transformer (ViT)

ViT tokenizes image patches and learns representations using self-attention. Performance is sensitive to training data volume.

C. GenConViT and DeepFake-Adapter

GenConViT integrates CNNs and transformers. DeepFake-Adapter fine-tunes a transformer for forgery detection, using residual attention blocks.

V. Dataset and Preprocessing

A subset of 200 images from the FaceForensics++ dataset (100 real, 100 fake) was used. Images were resized to 128×128 and normalized. Light augmentations (flip, scale) were applied to preserve artifacts.

VI. Methodology

A. Training Setup

All models were trained using binary cross-entropy loss and the Adam optimizer (LR = 0.001, batch size = 32). Early stopping was employed.

B. Ensemble Learning

A soft voting ensemble combined VGG16, GenConViT, and DeepFake-Adapter predictions. This helped stabilize predictions but didn't surpass VGG16 significantly.

C. Evaluation Metrics

Metrics used:

- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$
- F1-score: $2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$

VII. Results and Discussion

A. Confusion Matrices

CNNs like VGG16 identified more fakes than ViT. ViT underperformed due to lack of inductive bias and limited data.

B. Model Comparison

VGG16 achieved the best accuracy (70%). ViT scored 35%. The ensemble was stable but did not outperform VGG16.

C. Observations

ViTs require large-scale pretraining, while CNNs perform better with limited data. The ensemble method reduced prediction variance.

VIII. Future Work

Future improvements may include:

- Training hybrid CNN-ViT models on larger datasets
- Using multimodal data (audio + video)
- Applying explainability tools like Grad-CAM or attention heatmaps

IX. Conclusion

This project demonstrated the strengths of CNNs for DeepFake detection in small-data scenarios. Although ViTs offer theoretical advantages, they require substantial data or hybridization. Ensemble learning helped improve prediction consistency. Future work may involve deeper architectures, larger datasets, and explainability tools.

Acknowledgments

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

References

References

[1] A. Rössler 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. Available: https://scikit-learn.org

[5] Hugging Face Transformers. Available: https://huggingface.co/transformers/