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

Aditya Tejpal
Roll No: 102203330
Thapar Institute of Engineering and Technology
tejpaladitya@gmail.com

***Abstract*—The rapid rise of DeepFake content, powered by GANs and other generative models, has posed a major threat to digital media authenticity. This research evaluates the efficacy of pre-trained deep learning models, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), in binary classification of real and manipulated facial images. Four models—VGG16, ViT, GenConViT, and DeepFake-Adapter—are assessed using a subset of 200 images from the FaceForensics++ dataset. The highest accuracy of 70% was achieved using VGG16, while ViT struggled due to data limitations. Ensemble learning showed slight robustness improvement. The study highlights architectural trade-offs and presents directions for future enhancements such as multimodal learning and explainability tools.**

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

I. Introduction

DeepFakes are synthetic media generated by models like GANs, capable of mimicking human features with high fidelity. Their misuse can affect journalism, legal evidence, and public trust. This research investigates machine learning models, especially CNNs and ViTs, for their effectiveness in identifying DeepFake images. While CNNs are known for spatial locality and parameter efficiency, ViTs offer global feature modeling at the cost of larger data needs.

II. Background

A. Convolutional Neural Networks (CNNs)

CNNs are widely used in image classification due to their ability to extract spatial features via convolutional filters. They consist of convolutional layers, pooling layers, and fully connected layers.

B. Vision Transformers (ViTs)

ViTs adapt transformer architecture from NLP to vision tasks. They divide images into fixed-size patches and use self-attention mechanisms to model long-range dependencies. However, ViTs are data-hungry and lack the inductive bias inherent to CNNs.

III. Related Work

Previous studies have shown CNNs to be effective in detecting manipulated images [1], while ViTs show promise due to their capacity for modeling global context [2].

IV. Pre-trained Models

A. VGG16 (CNN)

A 16-layer CNN architecture popular for transfer learning in vision tasks.

B. Vision Transformer (ViT)

Processes image patches with self-attention. Struggles with small datasets.

C. GenConViT and DeepFake-Adapter

These hybrids integrate CNN-based spatial feature extraction and transformer-based global context modeling. DeepFake-Adapter includes residual attention modules tailored for forgery detection.

V. Dataset

A 200-image subset (100 real, 100 fake) from FaceForensics++ was used. Images were resized to 128×128 and normalized. Minimal augmentation was applied to retain DeepFake artifacts.

VI. Methodology

A. Preprocessing

High-resolution video frames were converted to static images. Augmentations included flipping and slight scaling.

B. Training Setup

All models used binary cross-entropy loss and the Adam optimizer with a learning rate of 0.001. Early stopping was employed with a patience of 10 epochs.

C. Ensemble Learning

A soft-voting ensemble of VGG16, GenConViT, and DeepFake-Adapter was created. While it improved stability, its performance was close to VGG16 alone.

D. Metrics

- **Accuracy**: $(TP + TN)/(TP + TN + FP + FN)$
- **Precision**: $TP/(TP + FP)$
- **Recall**: $TP/(TP + FN)$
- **F1-score**: Harmonic mean of Precision and Recall

VII. Results and Discussion

A. Performance

VGG16 outperformed all other models with 70% accuracy. ViT achieved only 35%, reaffirming the limitations of transformers in low-data environments.

B. Observations

CNNs utilize spatial hierarchies and inductive biases, making them efficient with small datasets. ViTs, though theoretically superior, require extensive data or hybridization.

C. Ensemble Evaluation

Ensemble learning slightly improved prediction consistency, but gains were marginal due to reliance on VGG16.

VIII. Future Work

Future improvements include:

- Training on larger, more diverse DeepFake datasets
- Incorporating multimodal inputs (audio + video)
- Using attention heatmaps and Grad-CAM for model explainability

IX. Conclusion

This project compared CNNs and ViTs for DeepFake detection using a small dataset. CNNs like VGG16 proved effective, while ViTs underperformed due to limited data. Ensemble learning provided some benefit but did not surpass the CNN baseline. The study provides insights into model selection, data sensitivity, and deployment considerations in real-world DeepFake detection systems.

Acknowledgments

The author thanks the Thapar Institute of Engineering and Technology for supporting this project.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.