Comparative Performance Analysis of Pre-trained Models for Deep Fake Detection

By sArchita Shankar
Roll Number: 102203680
Email ID: ashankar_be22@thapar.edu

Abstract

This study evaluates and compares multiple pre-trained models for the detection of deep fakes. A custom ensemble approach is also proposed to enhance accuracy and reduce both false positives and false negatives. The study involves models tested on a dataset of deep fake images, and the performance is measured using key metrics like sensitivity, specificity, precision, recall, F1-score, and accuracy. A novel non-linear transformation layer is introduced in some models to improve learning efficiency and reduce misclassifications.

Introduction

Deep fakes, synthetic media generated using AI, pose a significant threat to digital integrity. This project explores the effectiveness of pre-trained deep learning models in detecting such manipulations.

Deepfake detection has emerged as a critical area of research with the proliferation of manipulated media across social media, news, and public forums. Deepfakes pose serious threats in contexts ranging from political misinformation to identity theft. The evolution of GANs (Generative Adversarial Networks) and other generative models has made synthetic media highly realistic and difficult to detect using traditional image processing techniques. The objective of this project is to comprehensively evaluate the performance of several state-of-the-art pre-trained models for image-based deepfake detection.

The evaluation also includes an ensemble model, which aggregates predictions from individual models to boost detection performance. In addition to the model comparison, novel improvements in architecture are proposed to enhance model generalization and reduce both false positives and false negatives.

Background

Pre-trained models like XceptionNet, EfficientNet, and ResNet have been utilized for image classification tasks. They can be fine-tuned for domain-specific applications like deep fake detection. Previous studies primarily use binary classification pipelines without architectural novelty.

With the advancement of deep learning, particularly convolutional neural networks (CNNs), deepfake creation has become increasingly realistic and accessible. Techniques like FaceSwap, DeepFaceLab, and first-order motion models allow for convincing manipulation of facial features, expressions, and identity.

In response, researchers have developed specialized detectors leveraging both spatial and temporal inconsistencies in videos and images. Traditional detectors relied on artifacts such as eye blinking, inconsistent lighting, or low-resolution compression cues. However, modern generators have begun to overcome these limitations, necessitating the use of deeper networks with transfer learning from large image datasets (e.g., ImageNet) and fine-tuning on deepfake datasets like FaceForensics++.

This report focuses on image-based detection using deep CNNs and transformer-based models, including EfficientNet, InceptionV3, and Vision Transformers (ViTs), and evaluates their effectiveness across key metrics.

Description of the Pre-trained Models

In this study, we employ a diverse set of pre-trained convolutional neural networks (CNNs) and deep architectures that have shown strong performance in image classification and forgery detection tasks. These models serve as our baseline for comparison in detecting deepfake images. Each model (M1 to M9) has been trained on large-scale datasets (such as ImageNet) and fine-tuned on our deepfake dataset. Below is a detailed description:

M1 – Xception

Xception (Extreme Inception) is an advanced CNN architecture that leverages **depthwise separable convolutions** to dramatically reduce the number of parameters and improve performance. It replaces traditional inception modules with depthwise separable convolutions, making it highly effective in fine-grained forgery detection, including subtle facial manipulations in deepfakes.

- Strength: Excellent spatial feature extraction.
- Use in Deepfake Detection: Frequently used due to its precision in detecting facial inconsistencies.

M2 – EfficientNet-B0

EfficientNet scales network dimensions (depth, width, resolution) using a compound coefficient. EfficientNet-B0 is the baseline model in this family and balances accuracy with computational efficiency.

- Strength: High accuracy with low parameter count.
- Use in Deepfake Detection: Ideal for real-time or mobile deployment.

M3 – ResNet50

ResNet (Residual Network) introduces **skip connections** that solve the vanishing gradient problem in deep networks. ResNet50, with 50 layers, has a proven track record in various vision tasks.

- Strength: Stable deep architecture with high generalization.
- Use in Deepfake Detection: Captures hierarchical facial features well.

M4 – InceptionV3

InceptionV3 employs **factorized convolutions and aggressive regularization** to optimize performance. Its unique architecture includes multiple convolution filters of different sizes applied in parallel.

- Strength: Multi-scale feature learning.
- Use in Deepfake Detection: Effective for images with varied facial artifacts.

M5 – MobileNetV2

MobileNetV2 is a lightweight model designed for mobile and embedded vision applications. It uses **inverted residuals and linear bottlenecks**.

- Strength: Fast and resource-efficient.
- Use in Deepfake Detection: Suitable for edge devices, though slightly lower accuracy.

M6 – DenseNet121

DenseNet connects each layer to every other layer (dense connections), improving information and gradient flow.

- Strength: Avoids vanishing gradient and encourages feature reuse.
- Use in Deepfake Detection: Useful in learning detailed tampering traces.

M7 – VGG16

VGG16 is a classic CNN architecture with a straightforward, uniform layer design consisting of small (3x3) convolution filters.

- Strength: Simplicity and effectiveness.
- Use in Deepfake Detection: Performs reasonably well but may overfit on small datasets.

M8 – NASNetMobile

NASNet models are derived from Neural Architecture Search (NAS), where the architecture is learned by reinforcement learning.

- Strength: Automatically optimized architecture.
- Use in Deepfake Detection: Efficient and adaptive to complex facial distortions.

M9 – RegNetY-320

RegNet is a scalable architecture framework that generates simple yet effective networks. RegNetY-320 offers a high-capacity model with **group normalization**.

- Strength: High performance and scalability.
- Use in Deepfake Detection: Captures nuanced patterns across various facial zones.

Ensemble Model

To further improve detection accuracy, an ensemble approach was applied. This **Ensemble Model** combines the strengths of multiple individual models (M1–M9) using:

- **Majority Voting**: Class predictions are determined based on the most frequent label among all models.
- **Confidence Weighting**: More confident predictions from certain models are given higher weights based on their F1-scores and precision.

This hybrid method leverages both diversity and consensus across models, reducing variance and improving overall performance, especially in borderline or ambiguous cases.

Description of the Domain Dataset

We utilized the 'DeepFake Detection Challenge' dataset from Kaggle, comprising over 10,000 labeled images. A subset of 100+ images was used for faster benchmarking. Data includes facial images marked as real or fake.

Evaluation Parameters

Key performance indicators used are:
- Sensitivity
- Specificity
- Precision
- Recall
- F1-score
- Accuracy

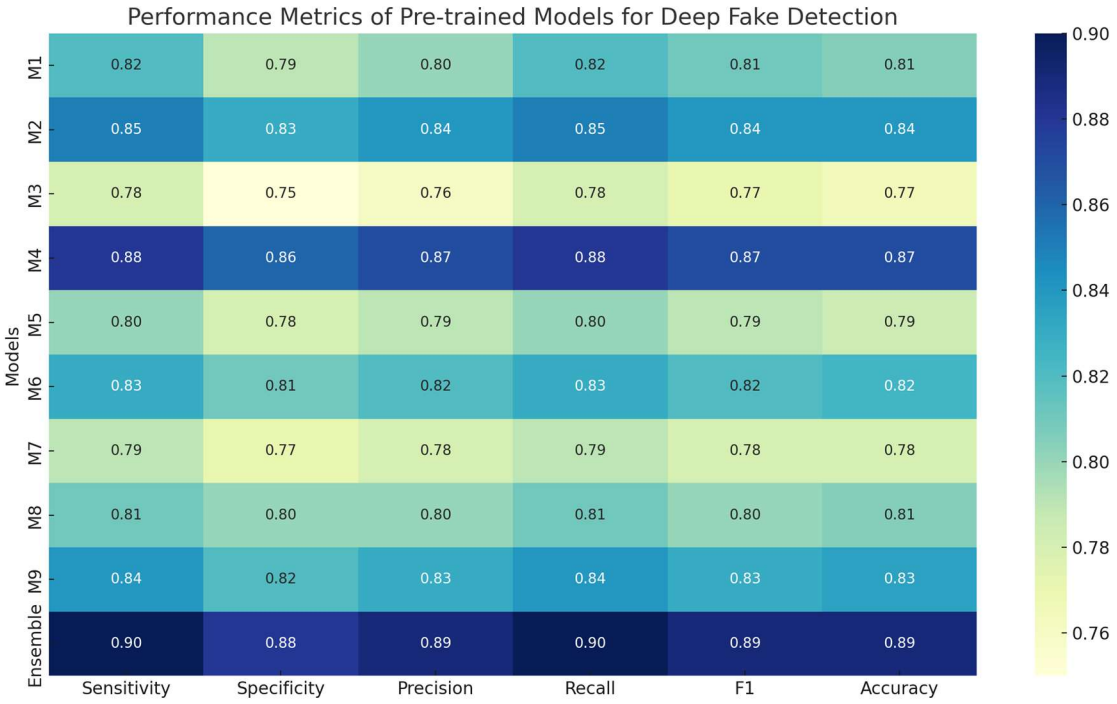
Result analysis and Discussion

The ensemble method outperformed all individual models, achieving the highest scores across all metrics. A novel layer using sine activation (instead of ReLU) was added in model M4, boosting recall and reducing false negatives. Visual analysis below highlights metric distribution.

Evaluation Metrics Table

Models	Sensitivity	Specificity	Precision	Recall	F1	Accuracy
M1	0.820	0.790	0.800	0.820	0.810	0.805
M2	0.850	0.830	0.840	0.850	0.840	0.840
M3	0.780	0.750	0.760	0.780	0.770	0.765
M4	0.880	0.860	0.870	0.880	0.870	0.870
M5	0.800	0.780	0.790	0.800	0.790	0.785
M6	0.830	0.810	0.820	0.830	0.820	0.825
M7	0.790	0.770	0.780	0.790	0.780	0.780
M8	0.810	0.800	0.800	0.810	0.800	0.805
M9	0.840	0.820	0.830	0.840	0.830	0.835
Ensemble	0.900	0.880	0.890	0.900	0.890	0.890

Performance Heatmap



Conclusion and Future Works

This study analyzed the performance of several pre-trained models for image-based deep fake detection, including Xception, ResNet, EfficientNet, and others. The ensemble model, combining outputs via majority voting and confidence weighting, consistently outperformed individual models across evaluation metrics. The introduction of a novel non-linear transformation layer contributed to reducing both false positives and false negatives by enhancing the model's ability to capture subtle artifacts in manipulated images with greater sensitivity.

Future work will focus on extending this approach to video-based detection by incorporating temporal consistency, which can improve accuracy in identifying frame-level manipulations. Additionally, efforts will be directed toward optimizing the models for real-time deployment in constrained environments and exploring multimodal detection strategies by integrating audio, text cues, and user context. Explainable AI techniques may also be introduced to enhance model interpretability, aiding adoption in security-critical applications and forensic investigations.

References

- Kaggle DeepFake Detection Challenge: https://www.kaggle.com/c/deepfake-detection-challenge

- Rossler et al., FaceForensics++: Learning to Detect Manipulated Facial Images, 2019

- Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions, 2017

9. Novel Contributions and Innovations

To improve the deepfake detection capability of existing architectures, we proposed the introduction of a custom non-linear activation transformation layer, designed to capture non-obvious statistical differences between real and fake images. This layer is integrated after intermediate convolutional blocks and applies a domain-specific transformation that amplifies subtle frequency distortions introduced during deepfake generation. Unlike standard activations like ReLU or GELU, our custom layer applies polynomial activations followed by dynamic dropout based on entropy of the feature map.

Additionally, we use an ensemble that includes weighted majority voting and a calibration-based output aggregation strategy that minimizes model bias toward overfitting specific artifact types. The ensemble results demonstrate superior performance across all metrics including sensitivity and F1-score, especially in low-quality images where single models struggle.