International Journal of Computer Techniques – Volume 12 Issue 3, May - June - 2025

Detecting Deepfakes Using Convolutional Neural Networks

Author: Nikhil Sharma
Department of Information Technology
Noida Institute of Engineering & Technology (AKTU), Greater Noida, India
Sharmanikhil.03n@gmail.com
Assistant Professor: Ram Kumar Sharma
Department of Information Technology
Noida Institute of Engineering & Technology (AKTU), Greater Noida, India
Ramsharma533@gmail.com

This research was conducted as part of the final year project in the Department of Information Technology at NIET, AKTU. Special thanks to the faculty for their guidance and to the open-source community for access to valuable datasets and tools. No external funding was received.

Abstract

Deepfakes are synthetic media generated using artificial intelligence, particularly deep learning techniques such as Generative Adversarial Networks (GANs). As these media become more realistic, the need for effective detection mechanisms has become critical. This paper presents a method for detecting deepfake videos and images using Convolutional Neural Networks (CNNs). A custom CNN model is proposed and evaluated on a publicly available deepfake dataset. Results demonstrate that CNNs can effectively learn spatial inconsistencies and subtle artifacts left by generative models. The paper also discusses performance metrics, comparative studies, and future directions in deepfake detection using deep learning.

Introduction

Deepfake technology, while impressive, poses significant risks including misinformation, fraud, and privacy invasion. These synthetically generated videos or images are often used maliciously to impersonate individuals, spread false information, or manipulate public opinion. Traditional forensic techniques are no longer sufficient due to the sophistication of generated content. Recent studies indicate that CNNs are well-suited for image classification tasks and can be adapted for deepfake detection by learning patterns that are hard for humans to perceive. This paper explores how CNNs can detect anomalies introduced by GAN-based deepfakes, aiming to develop a model that is both accurate and computationally efficient.

Literature Review and Problem Statement

The growing availability of deepfake creation tools has led to widespread misuse, necessitating the development of effective detection techniques. Several approaches have been explored:

- Biological Signal Analysis: Tracking inconsistencies like blinking or heartbeat.
- Frequency Domain Analysis: Spotting manipulation patterns in the frequency spectrum.
- Hybrid Models: Combining CNNs with RNNs or transformers.

Despite advancements, many existing models fail to generalize across different datasets and deepfake generation techniques. Moreover, lightweight and real-time deployable models are still lacking. This study aims to address these gaps by proposing a CNN-based solution that is accurate and efficient for real-world applications.

Methodology

Our approach involves the following steps:

- **Data Collection and Preprocessing:** Extracting frames from videos, resizing images to 128x128 pixels, normalization.
- **Model Design:** Constructing a CNN model with layers optimized for binary classification (real vs fake).
- **Training and Validation:** Using labeled datasets to train the model, monitoring loss and accuracy.
- **Evaluation:** Measuring performance with accuracy, precision, recall, F1-score, and ROC-AUC metrics.

Dataset

We use a subset of the DeepFake Detection Challenge (DFDC) dataset provided by Facebook. It includes:

- **Real Videos:** Unaltered content with original subjects.
- **Fake Videos:** Manipulated using various face-swapping and expression transfer techniques.

The dataset is split into 70% training, 15% validation, and 15% test data.

Preprocessing

Frames are extracted using OpenCV from each video at regular intervals. Each frame undergoes:

- **Resizing:** Adjusted to 128x128 resolution.
- **Normalization:** Pixel values scaled to [0,1].
- **Data Augmentation:** Applied random flipping, rotations, and zoom to enhance generalization.

Model Architecture

The Convolutional Neural Network (CNN) designed in this study is composed of several sequential layers tailored for image classification. It begins with three convolutional layers that extract local patterns such as edges and textures from input images. Each convolutional layer is followed by a max-pooling layer that reduces the spatial dimensions while retaining important features.

After the feature extraction phase, the data is flattened and passed through a fully connected dense layer, which helps the network learn non-linear combinations of the high-level features. To prevent overfitting, a dropout layer is introduced. The final layer uses a sigmoid activation function to output a binary classification — identifying whether the input image is real or fake.

This design balances accuracy and computational efficiency, making it practical for deployment in real-time or resource-constrained environments. The model is trained using the Adam optimizer and binary cross-entropy loss function, which are standard choices for binary image classification tasks.

Results

The model achieved the following performance on the validation dataset:

- **Validation Accuracy:** 87%
- **Precision:** 0.89
- **Recall:** 0.84

- **F1-Score:** 0.86
- **ROC-AUC:** 0.91

Table

Gesture Classification Accuracy per Sign

Accuracy of Responses

Metric	Value
Accuracy	87%
Precision	0.89
Recall	0.84
F1-Score	0.86
ROC-AUC	0.91

Discussion

The CNN model performs well in identifying deepfakes, especially on medium- to high-quality images. However, its performance degrades slightly with low-resolution or heavily compressed frames. The confusion matrix reveals that false negatives (real images misclassified as fake) were more common, which could be critical in real-world use-cases like social media moderation or legal evidence assessment. Enhancing robustness against video compression artifacts remains a key challenge. Despite these limitations, the model's overall performance shows promise for practical deepfake detection.

Conclusion

CNNs offer a promising solution for detecting deepfakes. With a relatively simple architecture, our model achieved strong performance on a challenging dataset. This suggests that spatial artifacts introduced by generative models can be effectively learned by CNNs. However, the model's limitations include sensitivity to image quality and lack of temporal context.

Future Scope

- Incorporating recurrent layers (LSTM, GRU) for temporal detection.
- Testing cross-dataset generalization.
- Using attention mechanisms to focus on key facial regions.
- Lightweight models for mobile deployment.
- Integrating face landmarks and eye-blink detection

References

- Afchar et al., MesoNet: a Compact Facial Video Forgery Detection Network.
- Rossler et al., FaceForensics++: Learning to Detect Manipulated Facial Images.

- Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions.
- Dolhansky et al., The Deepfake Detection Challenge Dataset. https://arxiv.org/abs/2006.07397
- Zhang et al., Detection of Deepfake Videos Using CNN-LSTM Networks.
- DFDC Dataset: https://ai.facebook.com/datasets/dfdc
- FaceForensics++ GitHub Repository: https://github.com/ondyari/FaceForensics
- DFDC Baseline GitHub (Facebook AI): https://github.com/facebook/DeepFakeDetection
- Deepfake Detection CNN+LSTM: https://github.com/danmohaha/deepfake-detection