Vision Transformer Based Deepfake Video Detection

Govinda Mandal[1] and Gautam Kumar[2]

[1] National Institute of Technology, Delhi, India
232211009@nitdelhi.ac.in
[2] National Institute of Technology, Delhi, India
gautam@nitdelhi.ac.in

Abstract. This paper presents a novel approach for detecting deepfake videos using Vision Transformers (ViTs) combined with Binary Cross Entropy (BCE) as the loss function during training and evaluation. With the proliferation of manipulated video content, effective detection methods are crucial for maintaining trust in digital media. Our method leverages the self-attention mechanism of ViTs to capture complex temporal dependencies across frames, enhancing feature extraction capabilities compared to traditional convolutional networks. We utilize two prominent datasets, FaceForensics and Celebs Deepfake, which provide a diverse range of manipulated video sequences for robust training and testing. The BCE loss function is employed to optimize model performance, particularly in binary classification tasks, by effectively measuring the difference between predicted probabilities and actual labels. Experimental results demonstrate that our approach achieves superior accuracy in identifying deepfake videos, highlighting the potential of ViTs in combating digital misinformation.

Keywords: Deepfake Detection, Vision Transformer (ViT), Binary Cross Entropy (BCE), Video Analysis, Computer Vision, Machine Learning, Adversarial Attacks, FaceForensics Dataset, Celebs Deepfake Dataset, Image Classification, Image Classification, Self-Attention Mechanism, Video Manipulation Detection, Neural Networks, Digital Media Forensics, Temporal Feature Extraction.

1 Introduction

Deepfakes represent manipulated media—be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread.

Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

In this paper, we present a novel deepfake detection system using ViTs, which outperforms conventional CNN-based approaches. Our work contributes to improving the detection accuracy and robustness of models against adversarial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.

2 Related Work

Deepfakes represent manipulated media—be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread.

Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

In this paper, we present a novel deepfake detection system using ViTs, which outperforms conventional CNN-based approaches. Our work contributes to improving the detection accuracy and robustness of models against adversarial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.

3 Objective

Deepfakes represent manipulated media—be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread.

Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

In this paper, we present a novel deepfake detection system using ViTs, which outperforms conventional CNN-based approaches. Our work contributes to improving the detection accuracy and robustness of models against adversarial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.

4 Algorithm

Deepfakes represent manipulated media—be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread.

Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

In this paper, we present a novel deepfake detection system using ViTs, which outperforms conventional CNN-based approaches. Our work contributes to improving the detection accuracy and robustness of models against adversarial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.

5 Dataset

Deepfakes represent manipulated media—be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread.

Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

In this paper, we present a novel deepfake detection system using ViTs, which outperforms conventional CNN-based approaches. Our work contributes to improving the detection accuracy and robustness of models against adversarial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.

6 Implementation

Deepfakes represent manipulated media—be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread.

Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

In this paper, we present a novel deepfake detection system using ViTs, which outperforms conventional CNN-based approaches. Our work contributes to improving the detection accuracy and robustness of models against adversarial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.

7 Results

Deepfakes represent manipulated media—be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread.

Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

In this paper, we present a novel deepfake detection system using ViTs, which outperforms conventional CNN-based approaches. Our work contributes to improving the detection accuracy and robustness of models against adversarial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.

8 Conclusion

Deepfakes represent manipulated media—be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread.

Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

In this paper, we present a novel deepfake detection system using ViTs, which outperforms conventional CNN-based approaches. Our work contributes to improving the detection accuracy and robustness of models against adversarial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.

Acknowledgment

Deepfakes represent manipulated media—be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread.

Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

In this paper, we present a novel deepfake detection system using ViTs, which outperforms conventional CNN-based approaches. Our work contributes to improving the detection accuracy and robustness of models against adversarial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.