Research Report on AI-Generated Image Detection using CNN Models and Vision Transformer

Authors:
Nilakhya Mandita Bordoloi, Abir Akash Baruah, Mrinmoy Shyam

Abstract

The proliferation of generative AI technologies has led to challenges in detecting fake images. This report addresses the risks associated with AI-generated images, emphasizing the threats of misinformation and identity fraud. It proposes a systematic solution, including understanding key concepts, preprocessing data, and experimenting with diverse models, aiming to contribute to a more secure online environment.

Date: January 9, 2024

MOTIVATION

Generative AI technologies make it easier to create fake identities, leading to more identity fraud cases. To tackle this problem, there's a need to improve online services for better security. By strengthening digital infrastructure and using strong authentication methods, the goal is to create a safer online environment that protects against the misuse of fake identities, ensuring a trustworthy digital space for everyone.

1 Introduction

1.1 The Dangers and Risk Involving AI Generated Images

Fake AI-generated images pose risks in various areas like social media and online identity verification. These convincing images created through advanced algorithms can lead to problems such as spreading false information, identity theft, and reduced trust in visual content.

A significant concern is the potential for misinformation and social manipulation, impacting the credibility of online content. In identity verification, these images can be exploited for creating fake identities, resulting in identity theft and fraudulent activities.

Deepfake technology adds to the worries, allowing the creation of realistic but fake videos or images. This can be misused for impersonation or creating fraudulent content.

To address these issues, there is an urgent need for AI-generated image detection mechanisms, often using advanced algorithms like deep learning models. These systems play a crucial role in identifying inconsistencies and unnatural patterns, helping online platforms, social media networks, and identity verification services strengthen their defenses against the impact of fake AI-generated images. This not only maintains the integrity and security of digital content but also protects against privacy infringements and fraud.

In essence, effective AI-generated image detection is essential for navigating the digital content landscape, minimizing the dangers associated with fake AI-generated images, and ensuring online spaces remain trustworthy, secure, and resilient.

1.2 Image Detection

Image detection in computer vision is a crucial process involving the use of algorithms and models to identify and locate objects or patterns within images. Key elements include preprocessing steps, Convolutional Neural Networks (CNNs) for feature extraction, object localization, and recognition. Transfer learning, with pre-trained models and fine-tuning, is commonly employed to boost efficiency. Detection methods may use region-based or anchor-based approaches, with post-processing steps to refine results. Recent advancements include Vision Transformer models that leverage attention mechanisms for capturing long-range dependencies in images.

In the context of identity verification, robust image detection plays a pivotal role in analyzing facial features, extracting biometric information, and validating identity documents, contributing to secure and reliable verification systems. Ongoing advancements in image detection techniques and deep learning models continue to enhance accuracy, efficiency, and adaptability across various computer vision applications.

2 Objective

With the advent of generative AI, it has become increasingly difficult to separate real data from AI-generated. The goal is to develop a model that can identify a fake photo created by AI.

3 Solution

The solution involves a comprehensive approach to address the problem in discussion. It begins by understanding key concepts related to the problem, focusing on CNN-based models like ResNet and VGGNet, as well as Transformer architectures, including Vanilla attention models and Vision Transformers. The process includes downloading the dataset, gaining familiarity with data types and attributes, and preprocessing the data through techniques like feature reduction. A predictive model is then employed to utilize combined feature knowledge for class prediction. The implementation phase integrates all elements into a pipeline, executed for a few epochs. The experimentation extends to using different models, combinations, and hyperparameter tuning. The analysis involves evaluating accuracy metrics with graphs, culminating in a conclusion that presents both qualitative and quantitative results obtained from the experiments. This structured methodology ensures a thorough exploration of the problem space and the potential effectiveness of various models.

4 Conclusion

In conclusion, this research report has delved into the challenges posed by AI-generated images, emphasizing the risks of misinformation and identity fraud. The proposed solution involves a systematic approach, encompassing the understanding of key concepts, preprocessing of data, and experimentation with diverse models. By focusing on CNN-based models,

such as ResNet and VGGNet, and incorporating Transformer architectures, including Vanilla attention models and Vision Transformers, the project aims to contribute to a more secure online environment.

The significance of robust image detection, as explored in the context of identity verification, underscores the need for continuous advancements in computer vision techniques. The systematic solution outlined in this report provides a structured methodology to address the complexities of AI-generated image detection.

By combining theoretical understanding with practical implementation, the report strives to pave the way for developing models that can effectively discern between real and AI-generated images. The ongoing efforts in this direction are crucial in ensuring the trustworthiness and security of digital content in an era dominated by generative AI technologies.