Detecting AI-Generated Images: A Simple Yet Effective Approach

Anonymous Author(s)
Affiliation
Address
email

Abstract

With the rise of AI-generated content, distinguishing between human-created and AI-generated images has become a critical challenge. We propose a simple yet effective approach using a lightweight CNN combined with handcrafted features such as edge detection and texture analysis. Our model efficiently classifies images while maintaining low computational cost.

1 Introduction

The ability to detect AI-generated content is a critical field of research, but detecting AI-generated images is especially important due to their many malicious applications. Generative AI models are capable of creating images realistic enough to be used to support the spread of misinformation online either through deepfakes or through other slanderous imagery that is often difficult to distinguish from real images. Beyond that, AI-generated images can pose ethical risks that can harm individuals and institutions due to their frequent use of copyrighted content or likenesses. Our work investigates ways to use modern machine learning models to determine the authenticity of images as efficiently and accurately as possible.

2 Related Work

As images usually do not come with any features, most AI image detection solutions involve deep learning models or convolutional neural networks (CNNs). Nayim et al. (2024) compared the performance of three different CNN models and found DenseNet was considerably more effective than the other models. However, their experiment was conducted using small images with resolutions no greater than 64x64. It is possible that a different model will be better suited for classifying the much larger images used in our experiment.

Deepfake video detection is a similar field of research that also frequently uses CNNs, but it can also incorporate a mix of other supervised and unsupervised learning techniques in an effort to reap the benefits of both approaches. A solution proposed by Soundarya and Gururaj (2025) uses the Dense Swin Transformer to perform spatio-temporal feature extraction to find "crucial clues" in inconsistencies across individual video frames.

Both AI image detection and deepfake detection share similar vulnerabilities. Their high computational costs render them impractical for deployment at a large scale. Certain image manipulations, such as Gaussian blurring, can make images considerably harder to classify correctly.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

3 Dataset and Preprocessing

The data for this project was provided by Women in AI through their Kaggle competition, and it includes an even mixture of real images from Shutterstock and AI images from DeepMedia. The whole dataset consists of 85,490 images. 79,950 of these images compose the training set, which is organized such that each human-made image corresponds to an AI image of the same subject. The training set is arranged in this way so that a classification model must base its decisions on something other than the contents of the image, since both the human-made and AI images will show the same content. Additionally, each image in the training set is labeled 0 for human-made or 1 for AI-generated. The test set consists of the remaining 5,540 images which are unlabeled.

3.1 Preprocessing Steps

- **Resizing:** All images are resized to 224×224 pixels.
- **Normalization:** Pixel values are scaled to the range $[0, 1]$.
- **Data Augmentation:** Horizontal flipping, random rotations, and brightness adjustments.

4 Feature Extraction

We enhance model performance using the following handcrafted features:

- **Edge Features:** Extracted using the **Canny** edge detector to highlight AI-generated artifacts.
- **Texture Analysis:** Applied **Histogram of Gradients** to capture texture inconsistencies and gradient change.

5 Methodology

Before testing more advanced models, we first established a baseline by testing a simple CNN.

6 Results and Discussion

Table 1 presents the classification performance.

Model	F1-Score
Baseline CNN	0.53

Table 1: Classification performance on AI vs. Human image dataset.

7 Conclusion

We demonstrate that a simple CNN combined with handcrafted features can effectively distinguish AI-generated images from human-created ones.

8 References

Anusha, Tenali, and A. Srinagesh. 2025. "Deepfake Video Detection: A Comprehensive Survey of Advanced Machine Learning and Deep Learning Techniques to Combat Synthetic Video Manipulation." 2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI), Multi-Agent Systems for Collaborative Intelligence (ICMSCI), 2025 International Conference On, January, 1033–41. doi:10.1109/ICMSCI62561.2025.10894187.

C, Soundarya B, and Gururaj H L. 2025. "Temporal Deepfake Detection Using CNN with Spatio-Temporal Features." 2025 17th International Conference on COMmunication Systems and NETworks

(COMSNETS), COMmunication Systems and NETworks (COMSNETS), 2025 17th International Conference On, January, 790–92. doi:10.1109/COMSNETS63942.2025.10885562.

Jiang, Justin. 2025. "Addressing Vulnerabilities in AI-Image Detection: Challenges and Proposed Solutions." 2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC), AI in Cybersecurity (ICAIC), 2025 IEEE 4th International Conference On, February, 1–9. doi:10.1109/ICAIC63015.2025.10849004.