Literature Review on Combating Deep Fakes: A Deep Learning Approach to AI-Generated Image Detection with Visual Reasoning

Madhav Krishnan u7735537

COMP4450

Table of Contents

1 Introduction

The recent rise of AI content has revolutionised various fields, leading to an unprecedented increase in AI-generated images. AI algorithms, particularly Generative Adversarial Networks (GANs), have made it possible to create highly realistic images that are often indistinguishable from real photographs. Deepfakes are AI generated media created with malicious intent and they have raised significant concerns due to their potential misuse in spreading misinformation, defamation, and propaganda.

As the realism of deep fakes continues to improve, the ability to discern these manipulated images from genuine content becomes increasingly challenging. This underscores the urgent need for effective detection methods to protect the integrity of visual content. This literature review is based on the research proposal "Combating Deep Fakes: A Deep Learning Approach to AI-Generated Image Detection with Visual Reasoning" where the goal is to not only detect AI-generated images but also to provide a visual explanation for the detection, thereby improving transparency and increasing understanding in the detection process.

This review examines the current state of AI image detection, explores various detection techniques, and identifies the limitations of existing models. It also aims to provide a comprehensive understanding of the field and highlight areas where further research is needed to develop better AI detection systems.

2 AI Image Detection

2.1 Necessity of AI Detectors

A study by Lu et al. (2024) titled "Seeing is not always believing: Benchmarking human and model perception of AI Generated images"[6] provides a comprehensive evaluation of both human discernment and contemporary AI algorithms in detecting fake images. The findings from the study show that humans believe more than one-third of the AI generated images are real compared to the AI misclassification rate of 13 percent. AI algorithms also don't have biases and this solidifies the superiority and clear need for AI detectors.

Despite the superior performance of AI detectors, the study also points out that existing AI algorithms still face significant challenges and need to be improved upon. We need for continuous improvement in AI detection models to further reduce the rate of misclassification and keep up with ever emerging techniques of image generation. Enhancing the accuracy and reliability of AI detectors is crucial for ensuring their effectiveness in real-world applications. Understanding how these models are trained can provide insights into areas where enhancements can be made, this will be further explored in the next section on the training and development of AI detection models.

2.2 Training and Development

AI image detectors are integral to the development and refinement of AI-generated image models. The relationship between AI detectors and image generators is symbiotic, creating an ongoing feedback loop that improves both systems. When developing an AI image generation model, detectors are employed to evaluate and enhance the realism of the generated images. This continuous cycle of improvement ensures that the generators produce highly realistic images, while the detectors become increasingly adept at identifying subtle signs of manipulation. A study by Bird and Lotfi (2024), titled "CIFAKE: Improving AI Image Generators and Detectors,"[2] underscores the importance of this feedback loop. The authors explain their model, which leverages convolutional neural networks (CNNs) to enhance both the generation and detection of AI-generated images. By using a feedback mechanism between the generator and detector, the model improves the realism of generated images while simultaneously increasing the accuracy of detection. The study also emphasizes the role of ensemble learning and adversarial training in improving the strength of AI detectors. Ensemble learning involves combining multiple models to enhance performance, while adversarial training exposes detectors to intentionally deceptive images, thereby improving their ability to identify manipulations.

Another significant contribution to this field is the work of Baraheem and Nguyen (2023), "Fine-Tuning CNNs for Robust Image Manipulation Detection."[1] This paper delves into techniques for fine-tuning CNNs on specific datasets to improve their detection capabilities. The authors highlight the benefits of adversarial training and ensemble learning, similar to the CIFAKE model,

and stress the importance of dataset specificity. By continuously training detectors on diverse and evolving datasets, the image manipulation detection is significantly enhanced, making these models more reliable in real-world applications.

By incorporating feedback loops, adversarial training, and ensemble learning, AI detection models can adapt to new challenges and maintain their effectiveness. This continuous improvement is essential as AI-generated image techniques evolve and become more sophisticated. Understanding and implementing these advanced training methods ensures that AI detectors remain at the forefront of detecting manipulated images.

Transitioning to an analysis of the existing AI image detection models, we can gain further insight into the effectiveness and limitations of current approaches.

2.3 Current Models Performance and Operation

The development of effective AI image detection models has been a crucial area of research, particularly due to the increasing sophistication of AI-generated images. Various models have been proposed and evaluated for their accuracy in detecting these images, each with its own strengths and weaknesses.

In a comparative study of AI image detection models, several tools were evaluated for their performance in identifying AI-generated images. One of the notable models, "CIFAKE,[2]" employs EfficientNets and Vision Transformers and achieved an impressive F1 score of 0.88 and an AUC of 0.95 on the DeepFake Detection Challenge dataset. The study also highlighted that convolutional and temporal techniques could achieve accuracy ranging from 66.26% to 91.21% on various synthetic data detection datasets. A key feature of these AI detectors is their ability to utilize chrominance components CbCr to detect minor pixel disparities indicative of synthetic images.

Bird and Lotfi (2024) [2]emphasize that most current AI detectors focus on providing a binary classification—real or fake—without explaining the rationale behind their decisions. This is a significant drawback because users, including researchers and practitioners, need to understand the reasoning behind the model's decisions to trust and effectively utilize these tools. The inability to pinpoint the exact features that the detector identifies as indicative of AI generation limits the transparency and interpretability of these models.

Moreover, the linear scale provided by the realism score offers limited scope for upgrading the models. Since it does not break down the contributing factors, developers find it challenging to fine-tune and enhance the model's accuracy based on specific weaknesses or errors in detection. This limitation necessitates a move towards more explainable AI models that can offer detailed insights and rationale for their classifications.

Our research proposal aims to address these shortcomings by integrating visual reasoning with AI detection. This approach not only seeks to improve the accuracy of detecting AI-generated images but also aims to provide clear explanations for the decisions made by the detector. By enhancing the interpretability and transparency of AI detectors, we can ensure more reliable and trustworthy

detection systems, paving the way for their broader adoption in real-world applications. In the next section let us look at detection techniques of AI detectors and the intricacies of image generation on which they operate.

3 Detection Techniques in AI Detectors

3.1 Passive vs. Active Detection

AI image detection techniques can be broadly classified into active and passive detection methods.

Active Detection: Active detection involves techniques that require interaction with the generative model. This approach can be highly effective when detailed information about the specific AI model is available, allowing for targeted analysis and detection. Active detection methods often probe the AI systems used to create images, identifying weaknesses or inconsistencies that can reveal the synthetic nature of the images. However, the creation of numerous new AI models daily makes it challenging for active detection methods to keep up. Without prior information about the model, active detection methods may struggle to provide accurate results. Despite these challenges, active detection can be highly useful and reliable for major and well-known AI models.

Passive Detection: In contrast, passive detection methods rely on analyzing images for inherent signs of manipulation without requiring additional information about the image generation process. These techniques identify artifacts, anomalies, or inconsistencies within the image itself. One significant advantage of passive detection is that it does not require prior knowledge of the image generation model, making it universally applicable to any image, regardless of the generative model used. Borji (2023), in "Qualitative Failures of Image Generation Models and Their Application in Detecting Deepfakes,"[3] underscores the importance of passive detection. Given the rapid proliferation of new AI models, passive detection is particularly useful as each new AI model introduces unique characteristics, making it impractical to rely solely on active detection methods.

Given these considerations, while active detection methods can be highly effective and more reliable for major and well-known AI models, passive detection methods are essential for the broad detection of AI-generated images. Passive detection methods leverage the inherent qualitative failures in AI-generated images, which will be explored in more detail in the next section.

3.2 Common Qualitative Failures in AI images

Qualitative failures in AI Generated images are common and can serve as key indicators for passive detection methods. These failures occur due to the inherent limitations of AI models in accurately replicating certain aspects of real-world images.

Common Qualitative Failures: Borji (2023)[3] outlines several qualitative failures that are particularly prevalent in AI-generated images. These include:

- **Hands and Faces:** AI models often struggle with generating realistic hands and faces, leading to unnatural appearances. These features are complex and highly detailed, making them difficult for AI to replicate accurately. Small errors in the shape, symmetry, or proportions of hands and faces can be easily noticed by humans.
- **Shadows and Lighting:** Inconsistencies in shadows and lighting can be a giveaway, as AI models may not accurately simulate how light interacts with objects. Real-world lighting follows complex physical rules, and deviations from these can make an image appear unnatural.
- **Scale and Geometry:** AI-generated images may exhibit errors in scale and geometric relationships, resulting in unrealistic compositions. Properly replicating the spatial relationships between objects requires a deep understanding of the real world, which AI models often lack.

These failures occur due to the inherent limitations of AI models in accurately replicating certain aspects of real-world images. Features such as hands, faces, shadows, and geometric relationships are particularly challenging for AI to generate because they involve intricate details and complex interactions that are difficult to model and learn.

Focusing on artifacts and blurs, this paper will look into specific types of anomalies that are common in AI-generated images. Artifacts and blurs are crucial indicators of image manipulation because they can highlight discrepancies that do not align with natural image creation processes. Unlike the detailed and often variable characteristics of hands or lighting, artifacts and blurs do not require prior training on specific features, making them more universally applicable for detection.

It is important to note that intentional blurs, such as those caused by AI-generated portrait effects with artistic blur, will have to be excluded from analysis in the dataset. Instead, the focus is on blurs and texture inconsistencies that indicate image manipulation and are easily verifiable by humans once highlighted. By concentrating on these anomalies, our proposal aims to enhance the effectiveness and reliability of passive detection methods for identifying AI-generated images. Let us look at algorithms we can utilize in images to detect these qualitative failiures in the next section

3.3 Artifact and Blur Detection Algorithms

To effectively detect AI Generated images, it is crucial to identify artifacts and blurs that signify image manipulation. Several algorithms are employed to detect these inconsistencies, each with its unique strengths and applications.

Texture Inconsistencies Fourier Transform: The Fourier Transform is a mathematical technique that transforms an image from the spatial domain to the frequency domain. This transformation makes it easier to analyze the frequency components of the image, which can reveal repeating patterns and textures that are not easily noticeable in the spatial domain. AI Generated images

often exhibit repetitive noise patterns and unnatural textures due to the limitations of generative models. By applying the Fourier Transform, we can detect these anomalies, which are indicative of synthetic content.

The Fourier Transform of an image $f(x, y)$ is given by:

$$F(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) e^{-i2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)}$$

Local Binary Patterns (LBP): Local Binary Patterns (LBP) is a texture descriptor that analyzes local patterns of pixel intensities in an image. It works by comparing each pixel with its neighboring pixels and encoding these comparisons into binary patterns. These patterns help identify texture inconsistencies that are often present in AI Generated images, such as unnatural smoothness or repetitive textures that do not occur naturally.

The LBP value for a pixel is computed as:

$$LBP(x, y) = \sum_{k=0}^{P-1} s(g_k - g_c) \cdot 2^k$$

where $s(x)$ is a thresholding function defined as:

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

and g_k is the gray value of the k-th neighbor, g_c is the gray value of the center pixel, and P is the number of neighbors.

Blurs and Edge Detection Sobel Operator: The Sobel Operator is an edge detection algorithm that computes the gradient of image intensity at each pixel, highlighting regions of high spatial frequency that correspond to edges. In the context of AI Generated images, the Sobel Operator can detect blurs that result from poor image synthesis. These blurs are often artifacts of the generation process and can indicate manipulation.

The Sobel Operator uses two 3x3 convolution kernels, G_x and G_y, to compute the gradient in the x and y directions:

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}, \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix}$$

The gradient magnitude is then computed as:

$$G = \sqrt{G_x^2 + G_y^2}$$

These algorithms collectively help in identifying the regions where artifacts are generated. By highlighting texture inconsistencies and blurs, they provide

valuable indicators that an image might be AI Generated. The combination of Fourier Transform, LBP, and Sobel Operator enables a comprehensive analysis of the image, ensuring that even subtle signs of manipulation are detected.

In our research proposal, we focus on these algorithms to enhance the detection process. By integrating them into AI detectors, we aim to improve the accuracy and reliability of identifying AI Generated images. The use of these techniques allows us to pinpoint specific anomalies, making the detection process more transparent and understandable for users.

4 Limitations of Current AI Detectors

4.1 Limited Binary Scaling

One of the significant limitations of current AI detectors is their reliance on simple binary scaling to indicate whether an image is AI-generated or not. As discussed previously, this binary classification provides a realism score but falls short in offering detailed insights into the specific reasons why an image was classified as AI-generated.

4.2 Difficulties in Detecting Partially Manipulated Images

AI-generated content often involves inpainting, where certain parts of an image are modified or replaced while leaving the rest of the image intact. Tools like Google's Magic Eraser [4] [5]and Samsung's object remover photo editor[8] enable users to remove objects from images seamlessly, making it challenging to detect these manipulations. In these cases, most of the image remains real, and only specific areas are altered, complicating the detection process.

Baraheem and Nguyen (2023)[1] discuss the challenges posed by inpainted images, noting that traditional detection methods may struggle to identify these subtle manipulations. This highlights the need for passive detection techniques that can identify and highlight specific anomalies within the image. By focusing on detecting texture inconsistencies and blurs, we can pinpoint areas that have likely been manipulated, even if the majority of the image appears real.

Relying on passive detection methods, as outlined in previous sections, becomes crucial in these scenarios. These methods do not depend on prior knowledge of the image generation model, making them more versatile and effective in identifying inpainted areas. Enhancing these techniques can help improve the detection of partially manipulated images, ensuring more comprehensive identification of AI-generated content.

4.3 Challenges with Grad-CAM for This Approach

Explainable AI (XAI) techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM), have been proposed to address the issue of limited interpretability in AI detectors. Grad-CAM generates visual explanations by high-

lighting the regions of an image that are most influential in the model's decision-making process. This can provide users with a clearer understanding of the detected anomalies and improve trust in the model's predictions.

The paper titled "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization"[7] introduces this novel technique, which uses the gradients of any target class flowing into the final convolutional layer to produce a coarse localization map. This map highlights important regions in the image, making it easier to understand why a model made a particular decision. The technique is class-discriminative and localizes relevant image regions for each class, enhancing the interpretability of model predictions.

Despite its advantages, Grad-CAM has limitations that make it less suitable for our approach. First, Grad-CAM is primarily designed for deep learning models with convolutional layers, which may not be present in all types of AI detectors used for image generation detection. Additionally, Grad-CAM provides a coarse localization map that may not capture the subtle and intricate details necessary for detecting minute artifacts and blurs indicative of AI-generated images. While Grad-CAM can highlight important regions, it might not be precise enough to identify specific texture inconsistencies or minor pixel disparities that are critical for accurate detection.

Furthermore, integrating Grad-CAM into our detection system would require significant computational resources and could slow down the detection process. This is a critical concern for real-time applications where speed and efficiency are paramount. For these reasons, while Grad-CAM is a valuable tool for understanding and interpreting deep learning models, it is not the optimal choice for enhancing the transparency and effectiveness of AI detectors in the context of detecting AI-generated images.

5 Conclusion

This literature review has highlighted the necessity and challenges associated with detecting AI-generated images, particularly in the context of combating deepfakes. The discussion began by emphasizing the superiority of AI detectors over human discernment. Despite their advantages, current AI detectors still face significant limitations, including the need for continuous improvement to keep up with evolving image generation techniques. The symbiotic relationship between AI detectors and image generators was explored, emphasizing the importance of feedback loops, adversarial training, and ensemble learning in enhancing the accuracy of AI detectors. Fine-tuning CNNs on specific datasets was also identified as a crucial factor in improving detection capabilities.

We differentiated between active and passive detection methods, noting that while active detection methods are effective for well-known AI models, passive detection methods are essential for the broad detection of AI-generated images. Qualitative failures in AI-generated images, such as inconsistencies in hands, faces, shadows, and geometric relationships, were identified as key indicators for passive detection. Specific algorithms for detecting artifacts and blurs, includ-

ing the Fourier Transform, Local Binary Patterns (LBP), and Sobel Operator, were discussed for their effectiveness in identifying regions with inconsistencies. The review also addressed key limitations of current AI detectors, such as limited binary scaling, difficulties in detecting partially manipulated images, and the challenges associated with integrating Grad-CAM for our approach. Our research proposal aims to enhance the robustness and transparency of AI detection systems by focusing on texture inconsistencies and blurs, laying the groundwork for future advancements in detecting AI-generated images.

6 Acknowledgements

I declare that this proposal is my own individual work, except where indicated otherwise. Any content derived from external sources have been appropriately acknowledged and referenced in accordance with academic conventions. I acknowledge the use of ChatGPT for grammar checking and language refining purposes, as well as latex code editing.

References

1. Baraheem, A., Nguyen, V.: Fine-Tuning CNNs for Robust Image Manipulation Detection. International Journal of Computer Vision (2023)
2. Bird, M., Lotfi, H.: CIFAKE: Improving AI Image Generators and Detectors. IEEE Transactions on Neural Networks and Learning Systems (2024)
3. Borji, A.: Qualitative Failures of Image Generation Models and Their Application in Detecting Deepfakes. Image and Vision Computing, ArXiv (2023)
4. Jarboe, N., Minnett, R., Constable, C., Koppers, A., Tauxe, L., Jonestrask, L.: Magnetics Information Consortium (MagIC) Database Interoperability Improvements: ORCID, EarthCube, Google, and PmagPy. In: IEEE Conference (2019)
5. Lin, G.E., Panigrahi, T., Womack, J., Ponda, D.J., Kotipalli, P., Starner, T.: Comparing Order Picking Guidance with Microsoft Hololens, Magic Leap, Google Glass XE and Paper. In: Workshop on Mobile Computing Systems and Applications (2021)
6. Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., Ouyang, W.: Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images. Neural Information Processing Systems (2023)
7. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision (2016)
8. Shi, M.: Deep Learning of Image Inpainting for Semiconductor Wafer Image Recognition. IEEE Conference (2022)