Combating Deepfakes: A Deep Learning Approach to AI-Generated Image Detection with Visual Reasoning

COMP4450 | Madhav Krishnan u7735537

1. Introduction

The digital landscape has been revolutionized by the emergence of powerful AI image generation tools like Midjourney and Stable Diffusion. These advancements offer unprecedented ease and accessibility for creating compelling visual content, impacting fields like graphic design, virtual reality environments, and even medical imaging. However, this progress comes with a downside: the proliferation of AI-generated images, particularly in the form of deepfakes.

Deepfakes are digitally manipulated images or videos created using AI algorithms, often with the intent to deceive or mislead viewers. The potential for misuse in spreading misinformation, defamation, and propaganda poses significant risks to trust in visual content. A recent study titled "Benchmarking Human and Model Perception of AI-Generated Images" [1] highlights this challenge, indicating that humans can only discern AI-generated images from real ones with 61.3% accuracy. This translates to 1/3rd of AI-generated images online potentially being mistaken for genuine content. This underlines the need for AI-driven solutions to address this growing concern.

The field of AI-based image generation detection has witnessed significant growth. Numerous models have been developed with varying degrees of effectiveness. The paper "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images" [2] discusses a model trained on the LAION-5B dataset, achieving a 92.98% accuracy in recognizing synthetic images. However, the effectiveness of these models can vary depending on the underlying generative techniques and noise implementations employed by different AI algorithms.

Furthermore, current approaches often classify images as real or fake without providing insights into the rationale behind the classification. This limits the potential for improvement and user comprehension. Identifying specific indicators, like impossible structures, artifacts, blurs, or evidence of specific generation techniques, is crucial. By pinpointing these anomalies, the system can assist in verifying the authenticity of visual content with a deeper understanding.

This would also aid with identifying AI-driven image editing tools, such as Google's "Magic Eraser" and Samsung's object eraser. These tools empower users to seamlessly remove unwanted elements from their photos, resulting in images where most of the content appears genuine and unaltered. Consequently, traditional detection methods may struggle to flag such images as manipulated due to their high degree of realism.

This research aims to develop an AI system that can not only detect potential tampering in images but also provide visual explanations for its assessments. **The focus** will be on a passive detection approach, prioritizing the identification of anomalies like artifacts, blurs, within existing images. **This scope** offers advantages in terms of broader applicability to various image sources and manipulation methods.

2. Background

2.1 Qualitative Failures in AI-Generated Images [3]: This study offers valuable insights into the inherent weaknesses of AI-generated images. By focusing on the concept of "qualitative failures," the research sheds light on specific areas where AI models struggle to produce realistic outputs. These vulnerabilities can be exploited during the detection process. For instance, the research might identify issues with AI-generated images in rendering fine details like hair, textures, or reflections. By training a deepfake detection system to recognize these specific qualitative failures, we can increase the accuracy of identifying generated content.

2.2 AI Detection of AI-Generated Images [4]: This research explores existing methods for leveraging AI models to detect deepfakes. It provides valuable context for our proposed approach, which aims to go beyond simple classification. The proposed approach will build upon the foundation laid by Baraheem and Tam's work, but with the crucial addition of explainable reasoning.

2.3 Artifacts and Exploiting Weaknesses[5]: This work delves into the vulnerabilities of Generative Adversarial Networks (GANs), a popular technique for creating deepfakes. GANs work by pitting two neural networks against each other: a generator that creates images, and a discriminator that tries to distinguish real images from the generated ones. By analysing the weaknesses within GANs, Zhang et al. identify specific artifacts and inconsistencies that can expose a deepfake's artificial origin. These "tells" can include issues with upscaling artifacts, unnatural blurring patterns, or inconsistencies in lighting and shadows. Our research will leverage the findings from this paper.

2.4 Detecting Photoshopped Images[6]: The work by Anand and Cao provides a valuable benchmark for comparison in the realm of image manipulation detection. Their research demonstrates the feasibility of using deep learning architectures, specifically ResNet, to detect manipulated images, particularly those edited using Photoshop software. While their work offers a strong foundation, it has a narrower scope compared to deepfakes. Deepfakes often employ more sophisticated techniques beyond simple photo editing tools, making them a more complex challenge for detection. The proposed approach will build upon the successes of this paper.

3. Research Questions

This research aims to address the following key questions:

RQ: Can an CNN be trained to effectively detect AI generation in images and give a response beyond a binary classification of real or fake?

Sub questions.

3.1 What specific features or anomalies within AI-generated images can a CNN be trained to recognize for effective detection? (e.g., artifacts, unnatural textures, inconsistencies)
3.2 How can the system be designed to not only detect AI generation but also provide informative explanations for its reasoning behind the classification?
3.3 How does the performance of the proposed CNN-based approach compare to existing methods for deepfake detection in terms of accuracy and efficiency?
3.4 What are the limitations of using a CNN architecture for this task, and how can these limitations be mitigated or addressed through alternative approaches?

The research questions as well as the sub question are explored below.

4. Methodology

4.1 Data Acquisition

The system utilizes the ArtiFact Dataset, a comprehensive collection of approximately 1.5 million AI-generated images derived from diverse GAN models and inpainting techniques. To ensure a representative sample for model training, stratified random sampling is employed to select a subset of 10,000 images. Images with intentional camera blur or those containing known AI signature patterns commonly detected by existing detectors are excluded. This focus is on identifying more subtle artifacts introduced during the AI generation process.

4.2 Detector AI Selection

The core detection component relies on a carefully chosen deep learning model (CNN) with a proven track record in image classification and anomaly detection. Pre-trained CNN models like VGG16 or ResNet50 will be explored and potentially fine-tuned on a specifically designed dataset containing real photographs and AI-generated images with labelled artifact regions. This training equips the model to identify and classify image regions exhibiting signs of manipulation during AI generation.

4.3 Anomaly Detection algorithms

- **Sobel Operator:** This classic edge detection algorithm identifies areas with unusual smoothness or blurriness within the images, which can be indicative of AI-generated artifacts.
- **Fourier Transform:** This mathematical transformation analyses the image's frequency domain. Repetitive patterns or unusual frequency components often signify artifacts introduced by AI generation algorithms. By analysing the image's frequency spectrum, the system can potentially uncover these hidden patterns.
- **Local Binary Patterns (LBP):** This technique analyses the local patterns of pixel intensities. Deviations from natural textures in LBP maps can suggest manipulation.

$$\mathbf{G}_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} * \mathbf{A}$$

$$\mathbf{G}_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * \mathbf{A}$$

Sobel Operator Credit: Wikipedia

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot e^{-2\pi i \left(\frac{ux}{M} + \frac{vy}{N}\right)}$$

Fourier Transform

$$LBP = \sum_{n=0}^{N-1} s(I_q - I_p) \cdot 2^n$$

4.4 Anomaly-based Removal and Weighted Realism Scoring

This stage leverages anomaly detection to simplify the image by removing suspicious areas and employs weighted realism scoring for the remaining content:

- **Full-Image Realism Score:** Utilize a pre-trained deep learning detector (CNN) like VGG16 or ResNet50 to obtain a baseline "realism score" for the entire image. This score reflects the detector's assessment of the image's naturalness.
- **Anomaly Detection and Removal:** Apply the chosen anomaly detection techniques (Sobel Operator, Fourier Transform, LBP) to identify potential tampering regions within the image. These regions might include areas with unusual smoothness, blurriness, repetitive textures, or other characteristics indicative of AI generation. Techniques like cropping or masking can be explored for removal.
- **Area-based Weightage:** Divide the remaining image (after anomaly removal) into small squares of varying sizes, ensuring complete coverage without overlap. Assign weights to each

square based on the proportion of the area it covers in the remaining image. Squares covering a larger portion will receive a higher weight.

- **Square Extraction and Scoring:** Extract these squares from the cropped image.
- **Weighted Realism Score:** Feed each square individually into the chosen deep learning detector for realism scoring. Multiply each square's realism score by its corresponding weight (from step 3). Sum the weighted products for all squares to obtain a weighted realism score.

4.5 Iterative Refinement for Maximum Realism Score

This stage focuses on refining the anomaly detection parameters and performing multiple iterations to achieve the maximum possible weighted realism score:

- Run the anomaly detection techniques (Sobel Operator, Fourier Transform, LBP) with various parameter settings.
- For each parameter combination, proceed to steps 2-5 of Anomaly-based Removal and Weighted Realism Scoring (see above).
- Select the parameter combination that results in the highest final weighted realism score (indicating the most effective anomaly removal).

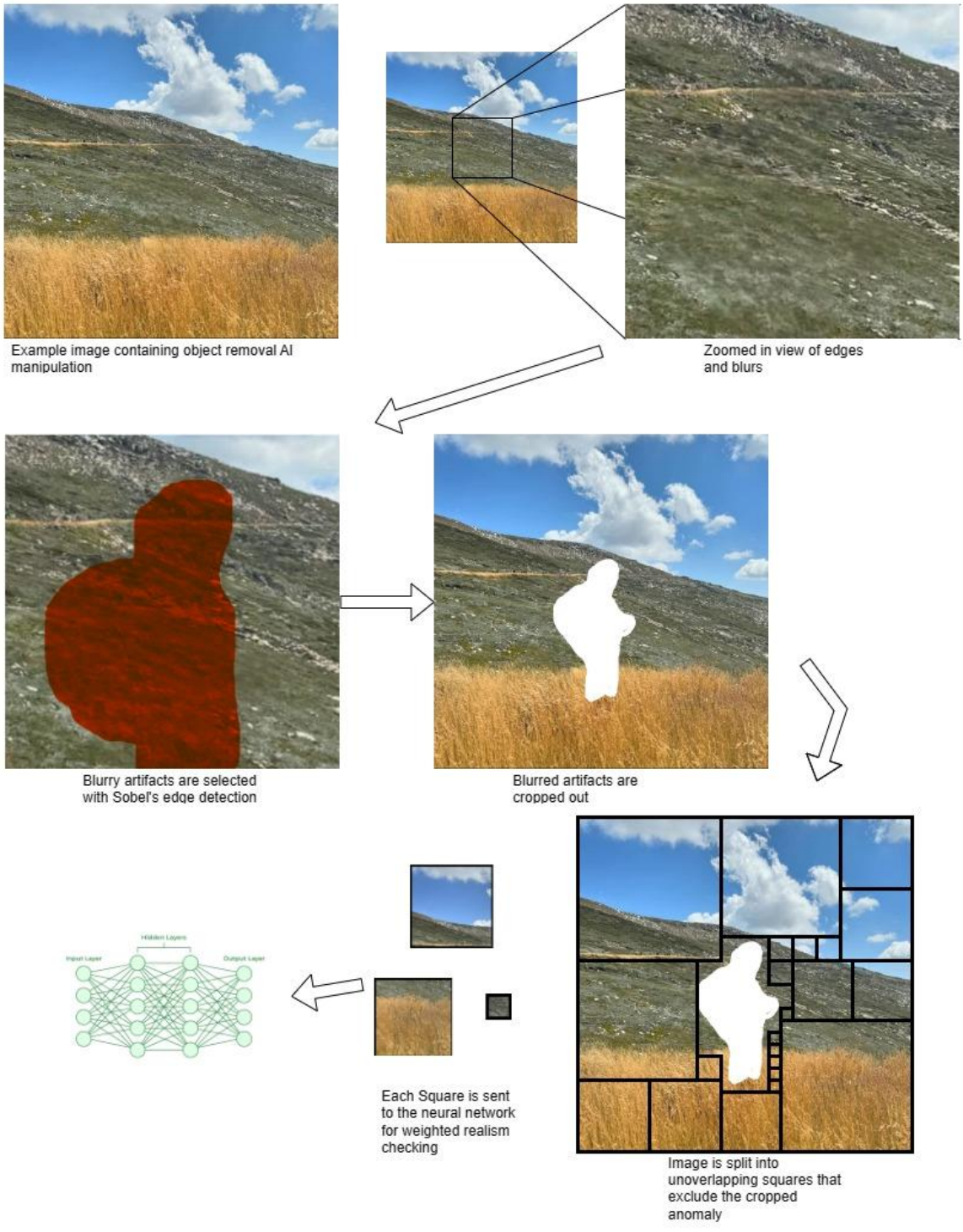
Monitoring Content Removal: Throughout the iterations, monitor the impact of anomaly removal on the realism score. This ensures you are not removing too much content that might affect the accuracy of the detector.

4.6 Decision Making based on Maximum Realism Score

Baseline vs. Maximum Realism Score: After the iterative refinement (step 4.4), compare the initial baseline realism score (obtained in 4.3) with the final maximum weighted realism score identified during the iterations.

Threshold for Improvement: Define a threshold for the minimum improvement in the realism score to determine the likelihood of AI tampering. Scores exceeding the threshold after anomaly removal suggest potential manipulation.

Visualization: Visualize the remaining image after anomaly removal using the parameter combination that achieved the maximum realism score.



An example of an AI object removed image that illustrates the Methodology steps of in 4.3

5. Evaluation Criteria

The proposed AI system for detecting tampering in images will be evaluated based on several key criteria:

Detection Accuracy:

- **True Positive Rate (TPR):** Measures the proportion of manipulated images correctly classified.
- **False Positive Rate (FPR):** Measures the proportion of real images incorrectly classified.
- **Area Under the ROC Curve (AUC):** A comprehensive metric that considers both TPR and FPR across various classification thresholds. Aim for a high AUC value closer to 1, indicating strong performance in distinguishing real from manipulated images.

Explanation Quality:

- **Relevance:** Do the visualized anomalies accurately correspond to the areas where manipulation is suspected?
- **Specificity:** Do the explanations highlight unique characteristics of AI-generated artifacts as opposed to natural image features?

Dataset Diversity: Test the system's performance on a variety of image datasets encompassing different AI generation techniques, manipulation methods, and real-world image types.

Noise Tolerance: Evaluate the system's robustness to noise and artifacts commonly present in real-world images (e.g., compression artifacts, camera noise).

Comparison with Existing Methods: Compare the detection accuracy, explanation quality, and efficiency of the proposed system with existing AI-based deepfake detection approaches.

F1 Score: Explore incorporating F1 score as a metric to assess the balance between precision and recall in identifying manipulated images.

6. Limitations:

6.1 Reliance on Deep Learning Detector Accuracy:

The system's effectiveness hinges on the chosen deep learning detector for realism scoring. If this detector struggles to distinguish real from manipulated images, the entire system's performance will suffer.

6.2 False Positives and Overaggressive Anomaly Detection:

Aggressive anomaly detection algorithms (Sobel Operator, Fourier Transform, LBP) might misinterpret natural image features as artifacts of manipulation, leading to false positives. This could result in:

- Removal of real image features crucial for accurate assessment.
- Reduced accuracy in identifying actual tampering.

6.3 Weight Selection and Square Division:

Dividing the image into small squares for weighted realism scoring introduces complexity in assigning appropriate weights to each square.

Uneven weighting could lead to:

- Overemphasis on certain image regions, potentially masking actual tampering.
- Underrepresentation of other regions, hindering the detection of subtle anomalies.

7. Research Significance

The proposed AI system, capable of detecting potential AI generation and visualising offers several advantages:

Enhanced Detection Accuracy: The multi-pronged approach to anomaly detection, combined with a well-trained detector AI, promises higher accuracy in identifying manipulated images compared to simpler classification methods.

Improved User Comprehension: Explanation functionalities provide valuable insights into the system's reasoning, increasing understanding among users.

Integration with AI Editing Tools: The core concepts of anomaly detection and visualisation can be integrated into existing image editing tools, empowering users to identify potential manipulations within their workflows.

Combatting Misinformation: By making it easier to detect and understand image manipulation, this research can contribute to the fight against misinformation and the spread of manipulated content online. It also sets the groundwork for detecting AI generation in images through other passive qualitative failure methods.

8. References

[1] Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X. and Ouyang, W., 2024. Seeing is not always believing: Benchmarking human and model perception of ai-generated images. Advances in Neural Information Processing Systems, 36.

[2] Bird, J.J. and Lotfi, A., 2024. Cifake: Image classification and explainable identification of ai-generated synthetic images. IEEE Access.

[3] Borji, A., 2023. Qualitative failures of image generation models and their application in detecting deepfakes. Image and Vision Computing, 137, p.104771.

[4] Baraheem, S.S. and Nguyen, T.V., 2023. AI vs. AI: Can AI Detect AI-Generated Images?. Journal of Imaging, 9(10), p.199.

[5] Zhang, X., Karaman, S. and Chang, S.F., Detecting and Simulating Artifacts in GAN Fake Images (Extended Version).

[6] Anand, K. and Cao, M., Doctored or Not: Detecting Photoshopped Images.

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 and Gemini for grammar checking and language refining purposes.