AI-Generated or Real

Manan Jain
mjain35@uic.edu
657477565

Hemanth Nagulapalli
hnagul2@uic.edu
670292806

Rishi Madhavarm
rmadhi4@uic.edu
654237930

Abstract

With the recent development of models that can generate large amounts of content, it has become paramount for people to be able to distinguish between what is handcrafted and what is generated. Technologies such as ChatGPT, Dall-E, SORA just to name the research done by OpenAI alone have pushed the limits of generated content such that it has become increasingly difficult to identify generated content. This project aims to conform a method to use Machine Learning to identify whether an image is AI generated.

I. Introduction

In recent years, artificial intelligence's capabilities in generating realistic images have reached unprecedented levels. Technologies such as Generative Adversarial Networks (GANs) and other deep learning models have created increasingly tricky images that are difficult to distinguish from those captured by traditional photography. This blurring line poses significant challenges and questions authenticity and trust in digital media.

The ability to differentiate between real and AI-generated images is not just a technical challenge, but a crucial need for a variety of applications. From content verification in social media to the integrity of journalistic practices and even legal evidence, the impact of this tool could be far-reaching. As AI-generated content becomes more pervasive, the potential for misuse grows, making it imperative to develop robust mechanisms to verify image authenticity.

The "AI Generated or Real" project was initiated against this backdrop, with the objective to create a reliable tool that could perform this differentiation with high accuracy. By focusing initially on a controlled dataset of apple images, chosen for their simplicity and uniformity, the project seeks not only to develop and refine this tool but also to establish a foundation for future research that could handle more complex and varied datasets.

This report outlines the development process of our classifier model, discusses its performance, including potential limitations such as sensitivity to lighting conditions, and explores the potential for its application in broader contexts. By advancing our understanding and capabilities in distinguishing between real and AI-generated images, we contribute to the ongoing discourse on digital authenticity and security.

II. Objectives

The "AI-Generated or Real" project aims to develop a classifier model that can accurately differentiate between authentic and AI-generated images. The specific objectives of the project are:

- **Develop a Classifier Model**: To design and implement a machine learning model that leverages state-of-the-art techniques to identify and classify images as real or AI-generated. The model should demonstrate high accuracy and efficiency in its predictions.
- **Achieve High Accuracy**: To achieve a high-accuracy classifier, the model must be reliable enough for potential practical applications and further research. We set a benchmark of 70% to establish a solid foundation for the model's performance.
- **Focus on a Controlled Dataset**: Initially, the model will be trained and tested on a dataset comprising images of apples. This focus will help control the variables and better understand the model's performance under simplified conditions.
- **Lay the Groundwork for Future Expansion**: While the current project is limited to Apple images, an objective is to establish a methodology and a baseline that can be extended to more complex and diverse image sets. This will allow future studies to adapt and scale the classifier to broader applications.

III. Literature Review

The capacity to distinguish between real and AI-generated images is a rapidly evolving area of research, underscored by the increasing sophistication of image synthesis technologies. This literature review, which forms the backbone of our project, highlights significant research that has influenced the development of our classifier, focusing on methodologies that enhance image analysis and authenticity verification.

A foundational piece of literature that informed our approach is “Direction-Aware Spatial Context Features for Shadow Detection and Removal” [1]. This study explores advanced techniques in detecting and analysing shadows in digital images, which are crucial for understanding the depth and authenticity of visual elements. The insights gained from this research have been instrumental in adapting similar context-aware approaches to distinguish between shadows and textures in actual versus AI-generated images, helping to enhance the accuracy of our classification model.

In addition to the specific methodologies derived from the above study, our literature review covers a broader range of works on Generative Adversarial Networks (GANs) and their applications in creating hyper-realistic images. These studies provide a deep understanding of the capabilities and limitations of current AI technologies in generating digital content, setting the stage for the necessity of practical detection tools.

Research on digital image forensics also plays a crucial role in our literature review. These studies focus on techniques for detecting manipulations and anomalies in digital images, closely related to identifying AI-generated content. By integrating these forensic techniques, we aim to enhance our classifier’s ability to perform nuanced analyses of image authenticity.

Overall, the literature underscores a growing need for reliable tools to assess the authenticity of digital images, driven by advancements in AI that challenge traditional verification methods. Our project is situated within this context, drawing on established research to build a solution that addresses these contemporary challenges.

IV. Methodology

The methodology for the "AI-Generated or Real" project was meticulously developed to focus on distinguishing real images from those generated by AI through specific inconsistencies that are often present in synthetic images. This section details our approach, the challenges encountered, and the limitations of our current model.

IV.i. Identification of Key Inconsistencies

Based on a literature review and preliminary analysis, we identified several typical inconsistencies in AI-generated images that could be exploited for classification. These include:

- **Lighting and Shadows**: AI-generated images often have anomalies in lighting and shadow, which do not conform to natural lighting laws.
- **Unnatural Texturing and Color**: These images may display textures and colors that are inconsistent with the expected properties of the objects they depict.
- **Background and Physical Placement Inconsistencies**: The arrangement of objects and backgrounds in AI-generated images can often lack the spatial coherence of natural images.
- **Unnaturally Plastic-Like Surfaces**: Synthetic images sometimes exhibit a glossy, plastic-like quality absent in natural environments.
- **Lack of Realistic Noise**: Digital noise occurs naturally in real photographs and is often missing or inconsistently applied in AI-generated images.
- **Image Specific Inconsistencies**: Certain details specific to the image's subject (like apples in our case) may not be accurately rendered in synthetic versions.

IV.ii. Focus on Lighting and Shadows

For this project, we focused primarily on analyzing lighting and shadows due to the depth of impact these elements have on an image's authenticity. This decision was informed by the significant research outlined in the literature review, particularly the insights from "Direction-Aware Spatial Context Features for Shadow Detection and Removal."

IV.iii. Data Collection

Our dataset comprises 800 real images and 800 AI-generated images of apples, organized into folders labeled "generated" or "real" accordingly. To consolidate so many generated images, we had to take many precautions:

- We used different image generation engines, such as Dall-E, Gemini, and Midjourney, that are available.
- We made sure to break the conversations so that the model does not follow a pattern of generated images.
- We tried out different prompts such that there are different types of images. For example, we used "Generate a realistic image of an apple," "Generate an image of apples," "Generate another image of an apple," and "Generate a different image of an apple."

For the real images, we took many photos and downloaded many images from the web, such as Pinterest.

If not already, all the images were cropped to a square shape and were downscaled to 192x192 pixels.

Additionally, we utilized the SBU Shadow dataset, which is instrumental in understanding image shadow characteristics.

IV.iv. Feature Engineering

We employed the Direction-Aware Spatial Context (DSC) approach from our literature review alongside the SBU Shadow dataset to develop a model capable of generating accurate shadow masks for any image. This model helps produce a mask that highlights shadows and another that indicates the light source. The light source mask shows either a single pixel indicating the direction of light (if the light source is not visible in the image) or the actual representation of the light source if it appears.

Our model then evaluates whether each image's light source and shadow are consistent. The model automates this consistency check, with each image tagging with a Boolean value representing whether the light and shadow elements are consistent (true) or not (consistent false).

IV.v. Model Training and Validation

The images were divided using an 80/20 split for training and validation purposes. We used a Convolutional Neural Network designed specifically for this task. The CNN architecture includes:

- Three convolutional layers with a 3x3 kernel, each followed by a 2x2 pooling layer to reduce spatial dimensions and enhance feature extraction.
- A flatten layer to convert the pooled feature maps into a single long vector.
- A dense layer with L2 regularization to prevent overfitting, followed by a dropout layer with a rate of 0.7 to further ensure the model generalizes well on unseen data.
- The final output layer utilizes a sigmoid activation function, classifying the image as '0' (generated) or '1' (real) based on the assessed consistency of lighting and shadows.

IV.vi. Testing

To perform testing, we took 20 new real images and 20 new generated images and then used the model to predict which image was real and which was generated.

IV.vii. Limitations and Scope

Due to time constraints and the complexity involved in integrating and implementing the full range of identified inconsistencies, this project iteration was limited to analyzing lighting and shadows. Future project versions will incorporate more initially identified inconsistencies, enhancing the model's robustness and applicability.

V. Results

The following graphs shows the results of the proposed methodology.

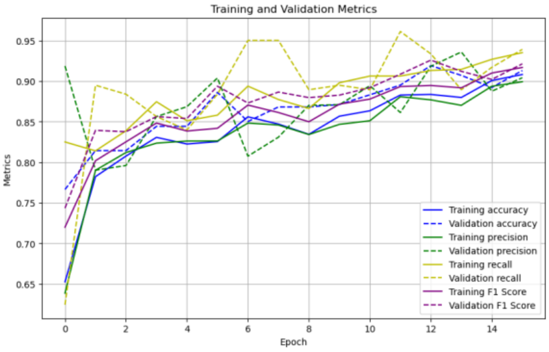

Figure 1 Training and Validation metrics

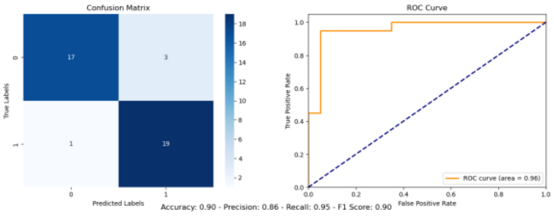

Figure 2 Testing metrics

We achieved training accuracy of about 90%, which was increased from about 70% if we used only the base classifier without the shadow analysis. There are some fluctuations in validation precision. However, that is also increasing over the epochs. The F1 scores are relatively stable and high, indicating a good performance on the training set. However, the validation F1 score still shows significant fluctuations, showing that the model is not able to generalize well.

From the Confusion matrix we can see that there is a very high accuracy of 90%, Precision of 86%, Recall of 95% and F1 score of 0.9. Also, from the ROC curve has an AUC of 0.96 which is very high.

VI. Discussion

The “AI-Generated or Real” project results reveal a high level of effectiveness in the classifier model’s ability to differentiate between real and AI-generated images. The testing results are particularly impressive, showcasing an accuracy of 90% and a confusion matrix that reflects excellent precision and recall. These results underscore the efficacy of our focused approach to shadow and lighting inconsistencies.

VI.i. High performance

While the testing results indicate excellent performance, integrating these with the validation metrics reveals a more nuanced picture. The validation results exhibit significant fluctuations, suggesting potential overfitting or the presence of challenging cases that the model struggles with. However, the overall trend across epochs shows a promising improvement, highlighting the model’s progressive refinement and instilling optimism about its future performance.

VI.ii. Challenges in Data Representation and Model Limitations

The diversity of the natural images in our dataset is extensive, yet the generated images present a spectrum of quality—from highly realistic to clearly artificial. This variance introduces complexity in model training, where highly realistic AI-generated images pose the most significant challenge, leading to the observed performance variability.

The differentiation in performance between clearly AI-generated and more nuanced synthetic images indicates that while the model excels in identifying overt inconsistencies, it may not uniformly handle subtle distinctions. This aspect, which we anticipated and are actively addressing, is crucial for understanding the model’s limitations and areas for improvement.

VI.iii. Impact of Shadow Analysis

Implementing shadow analysis has significantly improved the model’s performance. It has increased the base classifier’s accuracy from about 75% to approximately 90%. This substantial improvement not only validates our hypothesis but also reassures us about the effectiveness of our strategy in detecting AI-generated images. It also demonstrates the potential of targeted feature analysis in enhancing classification tasks.

VII. Conclusion and Future work

The "AI-Generated or Real" project has successfully developed a classifier model that distinguishes between real and AI-generated images with high accuracy. The model's focus on lighting and shadow inconsistencies has proven effective, achieving an accuracy of 90% in testing conditions, with a significant improvement from initial benchmarks.

VII.i. Key Achievements

High Accuracy: The model achieved a robust performance with a 90% accuracy rate, confirming its potential utility in practical applications.

Effective Use of Shadow Analysis: The model has demonstrated that these features are critical in distinguishing between real and AI-generated images by focusing on inconsistencies in shadows and lighting.

Automation and Scalability: The project has established a framework that can be scaled and adapted for broader applications, relying on automated processes to ensure consistency and efficiency.

VII.ii. Future Directions

While the project has successfully met its initial objectives, it also opens up several exciting avenues for further enhancement and exploration, showcasing its potential for continuous development and improvement.

Broader Image Diversity: Expanding the types of images and inconsistencies analysed by the model will be crucial. Including a more comprehensive array of subjects and more subtle AI-generated characteristics can enhance the model's robustness and generalizability.

Integration of Additional Features: Exploring other visual discrepancies, such as textural anomalies, unnatural color gradients, and digital noise, may provide deeper insights into image authenticity.

Advanced Computational Techniques: The current implementation utilizes several models to accomplish the task. After analyzing all the different features, we would like to consolidate all of this information using a single model that should be able to identify each feature and decide the type of image instead of relying on a pipeline that we have built up until now.

VII.iii. Theoretical and Practical Contributions

By providing a practical tool for distinguishing real from AI-generated images, this project makes a substantial contribution to the evolving field of digital image forensics. Its methodology and findings have far-reaching implications for areas where image authenticity is paramount, including media, cybersecurity, and digital content management.

References

[1] “Direction-Aware Spatial Context Features for Shadow Detection and Removal” IEEE Xplore - link

[2] “SBU shadow Dataset” Stonybrook University - link