Detecting AI Generated Images via Camera Sensor Patterns

Jun Lee
Computer Science, NYUAD
j10285@nyu.edu

Advised by: Yasir Zaki, Matteo Varvello

ABSTRACT

The past few years, after the introductions of many generative large language models such as OpenAI's Chat-GPT, the World Wide Web has seen a pattern of increasing AI generated contents including AI generated images. Naturally, as a reaction to this surge of AI generated content, AI generated content detection applications and APIs from companies such as Sight Engine and Hive Moderation have also been introduced. Majority of these detector products are based on machine learning models trained to identify contents generated by generative AI models such as Stable Diffusion or Dall-E models. Consequently, such detection products have edge cases in accurate detections such as contents altered by human after generation or content generated by AI models that wasn't included in the training data set.

This project aims to introduce a new method in AI generated content detection specifically in the area of photos. Firstly, a data set of edge cases for AI generated image detection will be gathered for this project's analysis. Then, this project will devise an algorithm that can identify AI generated photos from human taken photos using the photo-response non-uniformity (PRNU) noise which acts as a unique "fingerprint" for each camera device [6]. Using this algorithm, the project will further train a classification model that can detect AI generated photos including contents from the edges case data set. Various clustering and classification models have been tested for this project, and ultimately the Random Forest Classification model showed successful detection.

This report is submitted to NYUAD's capstone repository in fulfillment of NYUAD's Computer Science major graduation requirements.

جامعة نيويورك أبوظبي
NYU ABU DHABI

Capstone Seminar, Fall 2024, Abu Dhabi, UAE

KEYWORDS

Generative AI, Content Detection Applications, Edge Cases, Photo-Response Non-Uniformity, Cluster Model Training

Reference Format:
Jun Lee. 2024. Detecting AI Generated Images via Camera Sensor Patterns. In *NYUAD Capstone Seminar Reports, Fall 2024, Abu Dhabi, UAE.* 5 pages.

1 INTRODUCTION

In recent years, the World Wide Web has seen an unprecedented surge of AI-generated content after the introduction of popular large language models such as OpenAI's Chat-GPT. A notable component of this surge is AI-generated images, which are being increasingly incorporated into the digital landscape. Consequently, there has been a rapid development and deployment of tools designed to detect AI-generated content such as Hive Moderation's AI generated content detection or Sight Engine's AI image detection. A common foundational technology these tools use for their detection method is the implementation of machine learning models trained specifically to identify outputs from common generative AI models such as Stable Diffusion or Dall-E [4]. Despite their efficacy, these detection systems have limitations towards certain types of images such as content that has been manually altered post-generation or images produced by models not included in their initial training datasets.

Addressing these challenges, this project proposes a novel approach to enhance the detection of AI-generated images. The project is based on the use of photo-response non-uniformity (PRNU) noise, which serves as a unique "fingerprint" inherent in every camera device [6]. By utilizing this unique characteristic in image classification, the project aims to develop an algorithm capable of distinguishing AI-generated photos from those captured by humans. Using this new approach, this project could provide a more effective and comprehensive solution for identifying AI-generated content, enhancing the accuracy and scope of digital content verification. A potential use case of this capstone's product is the analysis of

the historic trend of AI generated content usage across the internet.

2 RELATED WORK

This section reviews several key studies and technological advancements relevant to our project's focus on detecting AI-generated images using PRNU noise patterns. T on the methodologies of current AI generated content detection tools were based on several papers such as David C. Epstein et al "Online Detection of AI Generated Images" [4] and Samah S. Baraheem et al. "AI vs. AI: Can AI Detect AI-Generated Images?" [2]. These papers provided insights on how machine learning models such as generative adversarial networks (GAN) or convolutional neural network (CNN) are trained and commonly used for AI generated content detectors.

The foundational work by Lukas et al. in their paper "Digital Camera Identification From Sensor Pattern Noise" [6] introduced the use of PRNU noise as a unique identifier for digital imaging sensors, establishing it as a crucial tool for camera identification in the field of digital forensics. Building on their previous paper, Jan Lukas explored the detection of image forgeries from sensor pattern noise in their paper "Detecting Digital Image Forgeries Using Sensor Pattern Noise" [5], demonstrating the practical applications of PRNU in various forensic scenarios. These papers were mainly referenced for the methodology of PRNU extraction. Additionally, this project utilizes the code developed by Dr. Miroslav Goljan [8] and his team, who are co-authors of the aforementioned papers, to implement the PRNU extraction process. The received code from Dr. Goljan's lab will be modified to attempt to extract PRNU from any uploaded image.

Furthermore, Guru Swaroop Bennabhaktula et al. "Source Camera Device Identification from Videos" [3] was referenced for extended use of PRNU analysis to video frames, enhancing the scope of device identification and forgery detection. Another paper referenced for the purpose of searching for alternative photographic property is Patel Manisha et al. "Beyond PRNU: Learning Robust Device-Specific Fingerprint for Source Camera Identification" [7] where the author introduces new "fingerprint" embedded in the low and mid-frequency bands of the image that can be extracted using convolutional neural network models. Altogether, these studies feature a strong potential of PRNU patterns as a novel approach to AI generated content detection.

Last but not least, Nan Zhong et al. "Rich and Poor Texture Contrast: A Simple Yet Effective Approach for AI-generated Image Detection" [9] was referenced for an alternative approach to novel AI generated image detection method utilizing inter-pixel correlations. Additionally, this project also used Nan's team's project's dataset of AI generated image and camera taken image to test and train the model.

This section will explore the technical background and significance of photo-response non-uniformity noise in digital imaging. Each camera has an imaging sensor that breaks down to several small pixels. These pixel sensors take photon inputs and convert them into digital signals. During this process of conversion, the voltage, before being converted to signals, passes through several filters such as the color filter array (CFA) which executes various color interpolation processes and algorithms [5]. Factors such as shot noise or sensor imperfections cause pattern noise [6]. The sensor patterns can be classified into two different types: Fixed Pattern Noise (FPN) and Photo Response Non-Uniformity noise (PRNU) [7]. For the purpose of identifying camera taken photos, this project will focus more on PRNU, which is primarily caused by different sensitivity of each pixel due to the inherent imperfection formed during the manufacturing process [6]. Due to this nature of causation, PRNU noise of each camera's photos can be treated as the "fingerprint" of the camera model or even the device itself. The core of this project is utilizing this characteristic property of an image taken by camera to identify human generated contents apart from the AI generated contents.

3 METHODOLOGY

This project utilized Dr.Goljan's PRNU extraction algorithm where a denoising formula using the Wiener filter is used. By subtracting the original image with the Weiner wavelet coefficients, the PRNU noise residual can be extracted.

$$n^k = p^k - F(p^k)$$

The PRNU extraction code was used for both datasets of both AI generated images and camera taken photos to train and test several classification machine learning models. Initially, clustering models such as K-means model and Gaussian distribution models were tried to group the AI generated images and camera taken photos for separation, but the individually extracted noises didn't show clear enough difference for the clustering models to separate the two. Therefore, in order to emphasize the difference in the PRNU patterns of AI generated images and camera taken photos, an additional step of rearranging the noise by centering the low frequency components using Fourier transform shift was applied. Since PRNU is primarily related to high frequency components, by applying the Fourier transform to feed, the resulting output should show clear difference in the center area of the extracting noise between AI generated images and camera taken photos if there is an inherent difference of the PRNU noise.

Also, instead of using clustering models, classification machine learning models such as convolutional neural network and random forest classification model was used.

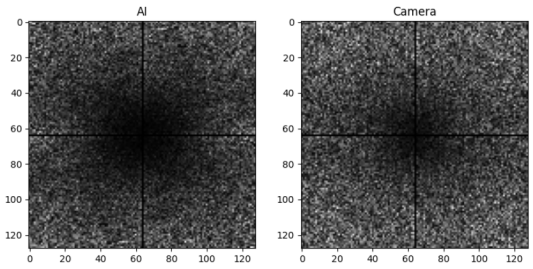

Figure 1: Sample image of Fourier Shifted PRNU.

3.1 Convolutional Neural Network

As mentioned previously, the common method for detecting AI generated images is through the use of trained convolutional neural network model.Despite the aformentioned shortcomings of the CNN model for AI image detections, the initial hypothesis was that if the model learned from the PRNU noise instead of the image itself, the inherent difference of the PRNU between AI generated image and camera taken photos can supplement the constraints of current AI detection methods.

The dataset for the AI-generated image for this experiment was acquired from Dr.Yasir's Communication Networks lab [1] , which includes 10,000 images generated by Stable Diffusion, Dalle 3, Dalle 2, and Midjourney. The dataset for camera taken photos was also acquired from Communication Networks lab where images from google map's reviews of restaurants were scraped. Each image was randomly cropped to 128x128 pixels before the PRNU extraction for the purpose of uniform data size, and Fourier transform was applied to the extracted noise as mentioned before. The model was trained on the normalized magnitude of frequency component of the image in 2D array format. The dataset was split 2:8 for valdiation and training. The model includes three convolutional layers with 32, 64, and 128 filters, each having a 3×3 kernel to optimize detecting small scale variations, ReLU activation, and L2 regularization with a coefficient of 0.001. Each convolutional layer is followed by a 2×2 max pooling layer and a dropout layer with rates of 0.25 to reduce overfitting. After the convolutional and pooling layers, the feature maps are flattened and passed through a dense layer with 128 neurons, ReLU activation, and L2 regularization. Another dropout layer with the rate of 0.5 precedes the output layer, which has a single neuron with a sigmoid activation function for binary classification. This was necessary since the PRNU difference between camera taken photos and AI genereted images was revealed small, and therefore overfitting was a potential issue for training the model. Data augmentation techniques, including random flips and rotations, are applied to the training data to further emphasize the difference between AI generated image and camera taken photos. The model is compiled with the Adam optimizer, binary cross-entropy loss, and accuracy as the evaluation metric and is trained for 10 epochs with a batch size of 32.

3.2 Random Forest Classification

An alternative method for classification model used was the Random Forest Classification model. An additional dataset of approximately 10,000 images, consisting of AI-generated images created using ProGAN, BigGAN, and StyleGAN models, as well as real images from Fudan University's lab, was acquired for CNN model testing. Three experiments were run using the Random Forest model, training on each of the aforementioned dataset of AI generated images. For this experiment, the images were randomly cropped to 256x256 pixels before the PRNU extraction for the purpose of uniform data size, and inverse Fourier transform was applied back to Fourier shifted noises and were flattened as a vector to train the random forest model. The dataset was split 2:8 for valdiation and training. The Random Forest classifier was initialized with 200 estimators, a maximum depth of 30, and a random state for reproducibility, balancing model complexity and generalization. However, after using the hyperparameter tuning via GridSearchCV, the configuration was set to 10 as maximum depth, 5 minimum number of samples required to split an internal node set, and 150 estimators. An acknowledged potential constraint is that for this experiment, the Random Forest model has been trained on generative AI models (ProGAN, BigGAN, StyleGAN), and therefore could be overfitted to each model.

4 RESULTS

4.1 Convolutional Neural Network

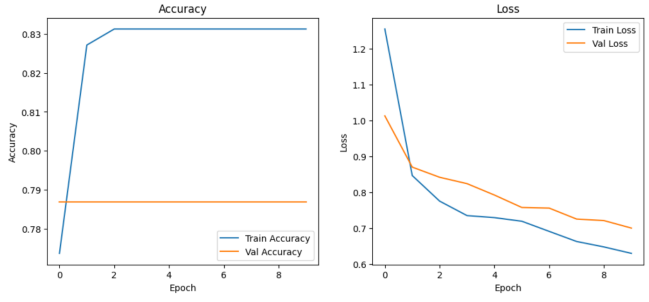

Figure 2: Accuracy and Loss Graph for CNN Model.

The results of the convolutional neural network revealed limited performance. The training accuracy rapidly rises to around 83% by the second epoch and plateaus for the remaining epochs. However, the validation accuracy remains stagnant at 79%, indicating that the model struggles to generalize to unseen data. This accuracy gap of approximately 4% suggests potential overfitting, where the model performs well on the training data but does not generalize effectively. Similarly, the loss trends show that the training loss decreases consistently from 1.25 to around 0.65, while the validation loss decreases at a slower rate, further reinforcing the model's difficulty in fitting the validation data. Despite using dropout layers (with up to 50% dropout) and regularization of L2 to mitigate overfitting, the generalization of the model remains limited.

4.2 Random Forest Classification

Table 1: Random Forest Results on ProGAN Dataset

Class	Precision	Recall	F1-Score	Support
0	0.94	0.88	0.91	3573
1	0.89	0.94	0.92	3629
Accuracy				0.91
Macro Avg	0.92	0.91	0.91	7202
Weighted Avg	0.92	0.91	0.91	7202

Table 2: Random Forest Results on StyleGAN Dataset

Class	Precision	Recall	F1-Score	Support
0	0.74	0.95	0.83	1100
1	0.93	0.68	0.78	1150
Accuracy				0.81
Macro Avg	0.83	0.81	0.81	2250
Weighted Avg	0.84	0.81	0.81	2250

Table 3: Random Forest Results on StyleGAN Dataset

Class	Precision	Recall	F1-Score	Support
0	0.74	0.95	0.83	1100
1	0.93	0.68	0.78	1150
Accuracy				0.81
Macro Avg	0.83	0.81	0.81	2250
Weighted Avg	0.84	0.81	0.81	2250

The results of the Random Forest model trained on ProGAN, BigGan, and StyleGan were somewhat satisfactory. The model performed best on the ProGAN dataset, achieving an accuracy of 91-92% with balanced precision, recall, and F1-scores for both classes. This indicates that the PRNU-based features extracted from ProGAN images are more distinguishable compared to real camera images, likely due to ProGAN's distinct noise patterns. For the StyleGAN dataset, the accuracy drops to 81%, with an imbalance between precision and recall: the classifier shows high precision for AI-generated images (0.93) but lower recall (0.68). This suggests that StyleGAN-generated images share more similarities with real camera images, making them harder to detect reliably. The BigGAN dataset shows the lowest performance, with an accuracy of 77% and lower precision-recall values, indicating that the PRNU noise patterns in BigGAN images are the least distinguishable.

5 CONCLUSION

From the results of the experiments so far, the Random Forest classifier model has shown higher performance over CNN model. A potential cause to this could be the lack of dataset variety and numbers for the CNN model to train effectively on since the Fourier shifted PRNU noise is inherently less diverse than raw images in terms of texture, edges, color, etc. Nevertheless, for both experiments, a more abundant and diverse dataset is required to further test the precision and accuracy of the models in identifying AI-generated images. The next steps for the spring semester will be training the models on a bigger, more diverse dataset that mixes all the AI generative models in order to avoid overfitting. Furthermore, additional feature vectors such as resolution or sharpness of the image will be extracted to further increase the accuracy of classifier models.

REFERENCES

[1] Nouar AlDahoul and Yasir Zaki. [n.d.]. NYUAD AI-generated Images Detector. https://huggingface.co/NYUAD-ComNets/NYUAD_AI-generated_images_detector

[2] Samah S Baraheem and Tam V Nguyen. 2023. AI vs. AI: Can AI Detect AI-Generated Images? *Journal of Imaging* 9, 10 (2023), 199.

[3] Guru Swaroop Bennabhaktula, Derrick Timmerman, Enrique Alegre, and George Azzopardi. 2022. Source camera device identification from videos. *SN Computer Science* 3, 4 (2022), 316.

[4] David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. 2023. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 382–392.

[5] Jan Lukáš, Jessica Fridrich, and Miroslav Goljan. 2006. Detecting digital image forgeries using sensor pattern noise. In *Security, steganography, and watermarking of multimedia contents VIII*, Vol. 6072. SPIE, 362–372.

[6] Jan Lukas, Jessica Fridrich, and Miroslav Goljan. 2006. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security* 1, 2 (2006), 205–214.

[7] Manisha, Chang-Tsun Li, Xufeng Lin, and Karunakar A Kotegar. 2022. Beyond prnu: Learning robust device-specific fingerprint for source camera identification. *Sensors* 22, 20 (2022), 7871.

[8] Morteza Darvish Morshedi Hosseini Miroslav Goljan. 2021. Camera Identification based on PRNU signal (Python implementation). https://dde.binghamton.edu/download/camera_fingerprint/. Version 1.1.

[9] Nan Zhong, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. 2023. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *arXiv preprint arXiv:2311.12397* (2023).