Fingerprint Extraction in Image Generative Models: Analysis of the frequency and spatial domain features

Research question

Which features—image residuals or noise residuals and phase or magnitude—are more effective for extracting fingerprints in the detection of AI-generated images?

Motivation

As image generative models progress dramatically and get far more higher investment than the synthesized images detection, the detection field struggles to keep pace with these advancements[1]. This problem emphasizes the importance of studying the features of generated images to determine the most reliable features in tracing the fingerprints produced by these models. Our research focuses on fingerprint extraction from AI-generated images, influenced by comparing noise residuals versus denoised image residuals in the frequency domain and spatial domain. Our goal is to identify the most reliable features for fingerprint extraction, which can then be utilized during the detection.

Approach

Data

We work with multiple fake images datasets and a real one, taking 1000 images from each dataset for the training. Frequency domain transformations are applied to these images to analyze and extract fingerprints from both noise and denoised image residuals.

BigGan	ArtiFact Dataset [2]
SD1.5	SD1.5 ImageNet [3]
SD2.1	Self-Generated
Flux	6k Flux Images [4]
Real Images	COCO Dataset [5]

Tools

The analysis is conducted using deep learning frameworks like PyTorch. Frequency-based residual extraction follows established methods from prior researches. For classification, we leverage advanced machine learning techniques.

Analysis Methods

- **Step 1: Residual Comparison**
  We begin by following the standard fingerprint extraction method as described in prior research [6], applying the same approach to both noise residuals and denoised image residuals in the spatial domain and to both phase and magnitude in the frequency domain. The goal is to determine which type of residual is more reliable in the extraction of the fingerprints.
- **Step 2: Determining the better feature**
  Knowing the fingerprints, we apply machine learning techniques to classify the images. Then we will evaluate our different fingerprints to determine which features were more reliable.
- **Step 3: Handling Multiple Generative Models**
  When the generative model that produced the image is unknown, we utilize fingerprints from multiple models for comparison, which identifies the closest match to determine the likely source.

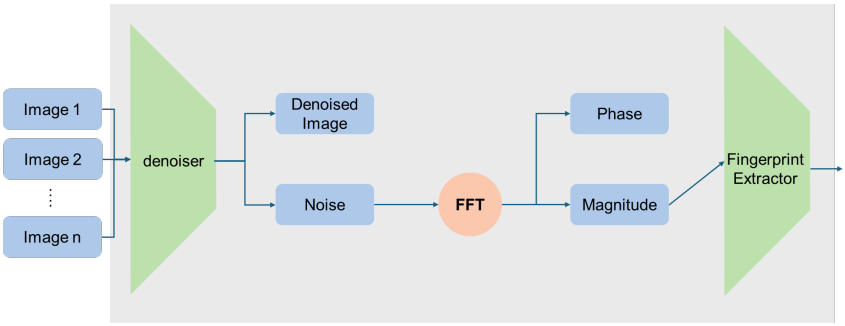


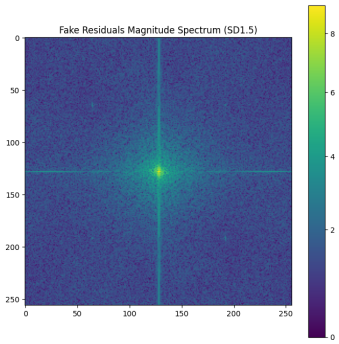
Figure (1) Pipeline of extracting the fingerprints using noise residuals' magnitude

Preliminary Results

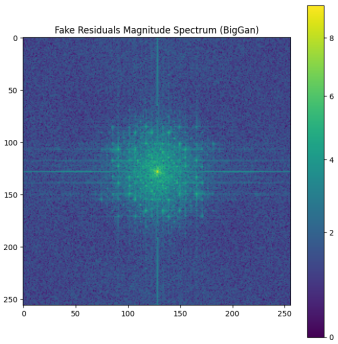
Metric	Denoised Image	Residual Image
Accuracy	0.52	0.85
Precision	0.88	0.97
Recall	0.04	0.72
F1 Score	0.07	0.83

Metric	Magnitude Denoised	Phase Denoised	Magnitude Residual	Phase Residual
Accuracy	0.76	0.56	0.83	0.49
Precision	0.79	0.56	0.82	0.43
Recall	0.71	0.48	0.84	0.10
F1 Score	0.74	0.52	0.83	0.16

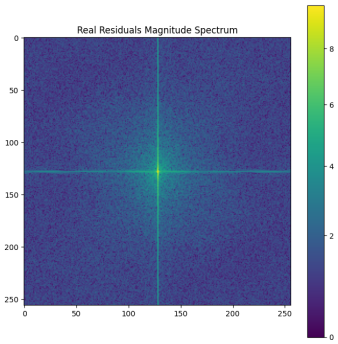
Here we compare between the noise residuals and denoised images in both spatial domain and frequency domain in the classification process.



(SD1.5) Magnitude of residuals



(BigGan) Magnitude of residuals



(Real) Magnitude of residuals

Here we compare between the magnitude spectrum of the fake images generated by SD1.5, BigGan and real images.We can spot notable differences between them which may lead to clearer fingerprints and easier classification. This is how we obtained these results:

Input Image : $I \in R^{3\times W\times C}$, Denoiser : D_n , Residuals in spatial domain : R

Getting the noise residuals using the difference between the original and denoised image

$$I - D_n(I) = R \tag{1}$$

Averaging the residuals

$$R_{avg} = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{2}$$

Taking the Fourier transform and plot the magnitude M

$$M = ||\mathcal{F}(R_{avg})|| \tag{3}$$

Results Discussion

1. Noise Residuals vs. Denoised Images: Our analysis shows that residuals, in both the spatial and frequency domains, are more reliable for fingerprint extraction than denoised images. Residuals consistently achieved higher performance across various tests.
2. Magnitude vs. Phase in the Frequency Domain: In the frequency domain, magnitude-based features outperformed the phase-based features in classification tasks. Magnitude features provided higher performance, making them more effective for fingerprint extraction.

Based on these findings, we conclude that our chances of success in the frequency domain are much higher, with magnitude holding the most valuable information for fingerprint extraction. Magnitude emerged as one of the most unique features, consistently yielding high accuracy during classification. Any progress we make in this problem would contribute directly to real life applications like filtering out fake data in medical related fields or in datasets that shall be used in training and fine tuning models. It worth noting that images used in training and testing processes didn't have any prior operations like compression or enhancement, but such operations might lead to significant loss in the fingerprint extraction process[7].

Previous Work

During the program, we worked on the spatial domain only and treated the whole image as our feature. We didn't try to split between the image and its noise residuals. We trained a classification model based on CNNs to see which parts of the image are more important in the classification process.

References

[1]Tariang, D., Corvi, R., Cozzolino, D., Poggi, G., Nagano, K., & Verdoliva, L. (2024, April 30). Synthetic image verification in the era of generative AI: what works and what isn't there yet. arXiv.org. https://arxiv.org/abs/2405.00196
[2]ArtiFact: Real and Fake image Dataset. (2023, February 23). Kaggle. https://www.kaggle.com/datasets/awsaf49/artifact-dataset
[3]Datasets at Hugging Face. (2001, June 16). https://huggingface.co/datasets/ek826/imagenet-gen-sd1.5
[4]Dataset with 6000+ FLUX.1 [dev] Images - 1024x768 and 768x1024 - v1.0 | Stable Diffusion Other | Civitai. (2024, August 8). https://civitai.com/models/631007?modelVersionId=705402 ek826/imagenet-gen-sd1.5 ·
[5]COCO - Common Objects in context. (n.d.). https://cocodataset.org/
[6]Deep image fingerprint: towards low budget synthetic image detection and model lineage analysis. (2024, January 3). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/10483843
[7]Dzanic, T., Shah, K., & Witherden, F. (2020). Fourier spectrum discrepancies in deep network generated images. https://proceedings.neurips.cc/paper/2020/hash/1f8d87e1161af68b81bace188a1ec624-Abstract.html