CSCI 1430 Final Project Report
Data Augmented AI Generated Detector

Everest Yang, Tanay Subramanian, Sujith Pakala, Sami Nourji
TA: Winston Li
Brown University

Abstract

This paper explores a Data-Augmented AI-Generated Image Detector to distinguish real images from AI-generated ones, addressing challenges posed by the rise of hyperrealistic content produced by generative AI. Using the CIFAKE dataset, we implement a CNN architecture with Fourier Transform features to evaluate their efficacy in identifying synthetic images. Our hypothesis is that incorporating frequency information via Fourier transforms, in addition to spatial domain information, into a CNN can enhance the detection of AI-generated images by leveraging frequency inconsistencies. This was validated by our research, as our best-performing baseline CNN achieved a testing accuracy of 96.92%, while our Fourier-based model reached an accuracy of 98.50%. Our findings highlight the potential of leveraging Fourier transforms for improved image classification, strengthening the growing field concerning digital authenticity.

1. Introduction

Misinformation and privacy are pressing concerns in today's modern world. As chatbots and generative AI become more sophisticated, such technologies can create hyperrealistic fake images that are nearly impossible for an individual to discern. Misusing these innovative technologies has significant implications for politics, social trust, and even individual security. For example, AI-generated images have already been involved in election interference, celebrity impersonations, and malicious pranks, underscoring the importance of a model that reliably detects fake images.

Consequently, our project becomes essential for verifying digital content's authenticity as realistic synthetic images can now be generated in seconds. However, the central challenge to solving this problem is that AI-generated content can replicate minute details such as lighting, shadows, and texture with high precision, making conventional detection methods less effective. Furthermore, it is not feasible to manually label AI-generated content at scale, highlighting the importance of automated tools in detecting artificial images.

We became familiar with these types of images through both social media and exploratory generation with tools such as Stable Diffusion. After lengthy discussions on the subject, we noticed that AI-generated images display textures that appear smoother than 'real images'. To test this theory, we decided to focus on AI-Generated image detectors, and introduce frequency domain information into the models through the Fourier Transform. This paper proposes a novel AI-Generated Image Detector to determine whether adding frequency domain information would enhance the model's ability to discern real images from AI-generated ones. Studies have shown that AI-generated images have unique characteristics - such as specific frequency patterns, smooth texture, and artifacts - which distinguish them from real images. By applying Fourier Transforms, we can quantify these differences in the frequency domain where smoothness and periodic artifacts are more evident. This approach aims to enhance the model's robustness in detecting artificial images.

2. Related Work

The rapid growth of generative adversarial networks (GANs) has created opportunities and ethical concerns regarding the misuse of synthetic images [3, 4, 2]. In recent academia, the CIFAKE dataset has become standard in distinguishing AI-generated images from real photographs. This dataset was created by generating synthetic images using latent diffusion to mirror the ten classes of the CIFAR-10 dataset. The synthetic dataset was paired with real images. Using a CNN, the study achieved 92.98% accuracy across 36 network topologies. Explainable AI techniques, using Gradient Class Activation Mapping, shows that the model focuses on small imperfections in the background instead of the main object [1]. We took inspiration from this dataset, learning from their technical implementation and results. Specifically, we explored their use of explainable AI techniques to refine our model's focus on relevant image features

such as edge sharpness and texture inconsistencies to improve interpretability.

Another research publication uses a lightweight method using CNNs with eight convolutional and two hidden layers to identify AI-generated images. It was tested on benchmark datasets and Sentinel-2 images, and outperforms four state-of-the-art methods with its lightweight architecture [1]. This inspired us to build our architecture in a similar way to achieve an equally high accuracy. We used similar batch normalization, density of convolutional layers, and dropout rates.

The main difference is that our paper explores adding a Fourier Transform to CNNs to enhance efficiency in image classification tasks. Based on their work, we integrated a Fourier Transform into our model to improve spatial-frequency feature extraction [5]. By leveraging the Fourier Transform, it accelerates training times by up to 71% with greater accuracy and reduced computational complexity. Given our limited hardware capabilities, we adopted their approach to balance accuracy and training time.

3. Method

TODO

4. Results

Model	**Test Accuracy**
Original CNN model [FIGURE 1]	98.58%
Fourier Transform only [FIGURE 2]	82.53%
Concatenated Fourier [FIGURE 3]:	
Baseline (Random Noise)	50.35%
Fourier Transform	95.36%
Combined Parallel Architectures [FIGURE 4]:	
Baseline	98.50%
Experimental conditions	98.50%

Table 1. Model Performance Comparison

Present the results of the changes. Include code snippets (just interesting things), figures (Figures **??** and **??**), and tables (Table **??**). Assess computational performance, accuracy performance, etc. Further, feel free to show screenshots, images; videos will have to be uploaded separately to Gradescope in a zip. Use whatever you need.

4.1. Technical Discussion

TODO

5. Conclusion

In this paper, we created and evaluated an AI-Generated Image Detector designed to distinguish real images from AI-generated ones. Our experiments demonstrated that incorporating Fourier features into the detection pipeline provided valuable insights, although the overall accuracy depended mostly on the CNN architecture. While our baseline model already achieved high accuracy, the addition of Fourier Transform features revealed weaknesses in specific scenarios, even though they are able to detect small differences in patterns and frequencies found in synthetic images.

A limitation of our project is the use of the CIFAKE dataset, which contains only 32×32 resolution images. Another limitation is that while this dataset is relatively new, it does not account for the significant advances in generative AI made in the last year. Hence, while this dataset allowed for efficient training and testing, it does not represent the variety or complexity of real-world images, limiting the generalizability of our findings. Future research should prioritize datasets with higher resolutions and more diverse content to ensure better performance in real world images/applications. As generative AI continues to evolve, solutions like the one proposed in this paper will play an important role in preserving digital authenticity and societal trust.

References

[1] J. J. Bird and A. Lotfi. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *IEEE Access*, 2024. 1

[2] Luca Romeo, Federico Federico, Anxhelo Dervishi, Raoul M. M., and Luigi C. Faster Than Lies: Real-Time Deepfake Detection using Binary Neural Networks. *arXiv preprint*, arXiv:2406.04932, 2024. 1

[3] Victor L. Thing. Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers. In *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 246–253. IEEE, July 2023. 1

[4] Yong Wang, Yifan Hao, and Ai Xia Cong. Harnessing Machine Learning for Discerning AI-Generated Synthetic Images. *arXiv preprint*, arXiv:2401.07358, 2024. 1

[5] Jakub Zak, Anna Korzynska, Antonina Pater, and Lukasz Roszkowiak. Fourier Transform Layer: A Proof of Work in Different Training Scenarios. *Applied Soft Computing*, 145:110607, September 2023. 2

Appendix

Team contributions

Everest Yang: I developed the code for integrating the Fourier Transform into our model, including the implementation for concatenating the flattened vector dimensions. Additionally, I conducted an extensive review of related works and research papers to identify optimal CNN architectures. I also contributed significantly to the writing of the paper, specifically on the Introduction, Related Works, Technical Discussion, Conclusion,

and key parts of the Method section. Furthermore, I formatted the entire document into a comprehensible research paper format using LaTeX and a .bib file for our references.

Tanay Subramanian: I was responsible for finding literature concerning current research concerning neural networks and the integration of Fourier transforms into classification tasks. Additionally, I helped develop the code for the baseline model architecture, in addition to working on the final paper, contributing to the Abstract, Introduction, Methods, and Conclusion sections. I also worked extensively on the final presentation slides, helping summarize our research in a concise and visually appealing manner.

Sujith Pakala: Once we realized that our project would require more computing power than our computers or Google Colab would allow, I took on the role of understanding and debugging our set-up with Oscar to be able to test our models in a time efficient manner. I also spearheaded the development of the code for our baseline without the fourier transform and the code for just using the fourier transformation after investigating architectures used historically in the literature. I also worked with Sami to design our "experiments" and created the graphs used in the report.

Sami Nourji: I took on the role of Project Manager for this final project, helping develop the project idea, and managing the allocation of tasks among teammates. My work within the project involved developing the Fourier experiments, model architectures, and training the model. I also worked on interpreting the model results and formulating the write up conclusion