Fusing Spatial and Frequency Data for Robust Detection of DALL-E 3 Generated Images

Justin Cargiulo
University of Maryland Electrical and Computer Engineering Department
College Park, Maryland, United States
jcargiul@terpmail.umd.edu

***Abstract*—Artificial intelligence has demonstrated remarkable capabilities in generating content that is nearly indistinguishable from content created by humans. These capabilities span across many mediums including, writing, videos, and images. It is more important now than ever to have techniques to distinguish AI-generated images from human-generated images. This research aims to identify signatures in images generated by OpenAI's DALLE-3 by comparing the classification accuracy of a model trained on images versus a model trained on both images and their 2-dimensional Fast Fourier Transforms. Two datasets were used for training, a collection of human-created content (pictures, artwork, etc.) and a collection of images generated by DALL-E 3, both consisting of 3,000 images. Both the image-only model and the hybrid model were trained on these datasets using Convolutional Neural Networks with nearly identical architectures, and the performance metric used to compare them was validation accuracy. After training, it was found that the hybrid model was 12.0% more accurate in distinguishing DALL-E 3-generated content from human-created content than the image-only model. These findings suggest that the magnitude spectra of images provide valuable information to a model, making it more robust when detecting AI-generated images. Further research is required to test the generalization of the results to other reverse diffusion models such as Stable Diffusion and other mediums such as video.**

***Index Terms*—Machine Learning, Artificial Intelligence, Reverse Diffusion, Fast-Fourier Transform, AI-Generated Images**

I. Introduction

As Artificial Intelligence improves, the threat of AI-generated images and deep fakes will continue to grow. Distinguishing between what content is created by AI and what content is created by humans is more important than ever.

The current standard method for AI image generation models is the Reverse Diffusion method (fig. 1). This method, used by the most popular models such as DALL-E 3 and Stable Diffusion, is trained by iteratively adding noise to billions of images and studying the change over every iteration [1] [2]. Once trained, the model then takes a prompt and removes noise from an image of only noise until the final output image is left.

There are currently many efforts to develop consistent and accurate methods to distinguish between AI images and human images. In a recent study conducted at Columbia University, Xu Zhang, Svebor Karaman, and Shih-Fu Chang identified significant patterns in the frequency domain of AI-generated images and trained a model on only spectra as an input. They determined that a model could be trained using only the magnitude spectra of the original image as an input to the model and obtain better detection accuracy than when using only the original image.

The goal of my research was to extend the current, novel work being done in AI content detection using the frequency domain. To do so, I wanted to determine if a hybrid model trained on the image and its magnitude spectra would produce higher accuracy results than training the model on just the image. No model has been researched using this hybrid approach, and limited research has been conducted on the viability of spectra-trained models.

The magnitude spectrum is computed using a 2-dimensional Fast Fourier Transform of an image. The 2D Fast Fourier Transform (FFT) of an image is mathematically represented as:

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) e^{-i2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)}$$

where $F(u,v)$ is the Fourier transform of the image, $f(x,y)$ is the intensity of the pixel at position (x,y), and M and N are the dimensions of the image.

Both models will be using a Convolutional Neural Network. A Convolutional Neural Network (CNN) is a deep learning algorithm used commonly for image classification purposes. Convolutional Neural Networks adaptively train from low-level features such as edges or colors to high-level features such as objects and shapes. Convolutional Neural Networks use convolutional layers to create feature maps by applying certain filters. The equation for these convolutional layers on images is mathematically represented as:

$$y_k = \sum_{c=1}^{C} (x_c * w_{k,c}) + b_k$$

where:

- y_k is the output feature map from the k-th filter.
- x_c is the c-th channel of the input data.
- $w_{k,c}$ represents the weights of the k-th filter for the c-th channel.
- b_k is the bias for the k-th filter.
- C is the number of input channels.

- $*$ denotes the convolution operation.

If the hybrid model is more robust and produces higher accuracy results than the image-only model, this would indicate the possibility of digital signatures in AI-generated images, invisible to the human eye, that can be used for detection purposes. This capability could provide further insight into the relationship between the reverse diffusion process and the frequency domain and could promote more ethical and responsible AI use.

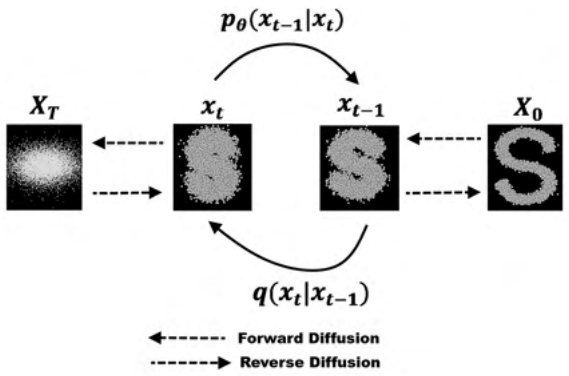

Fig. 1. The reverse diffusion process, from TowardsDataScience.com [3]

II. Methodology

Both the baseline, image-only trained model and the hybrid model were trained on a dataset of 3,000 DALL-E generated images and 3,000 human-created photos and art obtained from the "DALL-E Recognition Dataset" authored by Gaurav Dutta on Kaggle.com. Both classes of images are collections of diverse images and artwork, depicting humans, nature, and fictional characters, to name a few. Sample images from each class can be seen in Figure 2 and Figure 3.

Fig. 2. Sample images from the human-generated content dataset

Fig. 3. Sample images from the DALL-E 3 generated content dataset

A. Image Data Generators

The image-only and hybrid models required image data generators to preprocess the images.

1) Image-Only Model Image Data Generator: For the image-only model, each pixel value was rescaled by 1/255 to transform the range from [0, 255] to [0, 1], which helps the neural network learn more efficiently as normalized values are generally easier to process. There were also certain data augmentation strategies used to combat over-fitting and make the model more robust. These data augmentation techniques include random rotations, horizontal flips, and image shearing. These preprocessing techniques are used on both datasets for the model.

2) Hybrid Model Image Data Generator: The hybrid model image data generator also normalizes the original image by rescaling it by 1/255. The generator then converts the images to grayscale to simplify computing the Fast Fourier Transform. The Fast Fourier Transform was computed using OpenCV's cv2 library. The magnitude spectrum produced by the FFT is then logarithmically scaled to enhance the visibility of the spectrum's details. The image data generator then concatenates the magnitude spectrum to the original, normalized, RGB image as an extra, fourth layer. Images from the human-generated content dataset and their associated FFTs can be seen in Figure 4, and images from the DALL-E 3 generated content dataset and their associated FFTs can be seen in Figure 5.

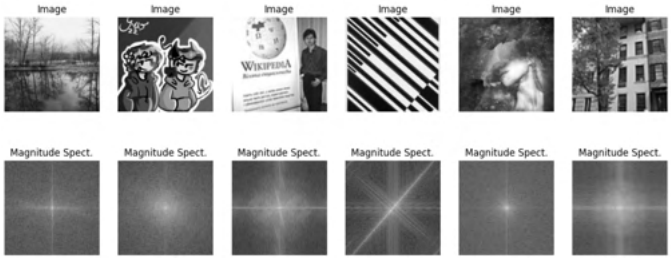

Fig. 4. Sample images and their FFTs from the human-generated content dataset

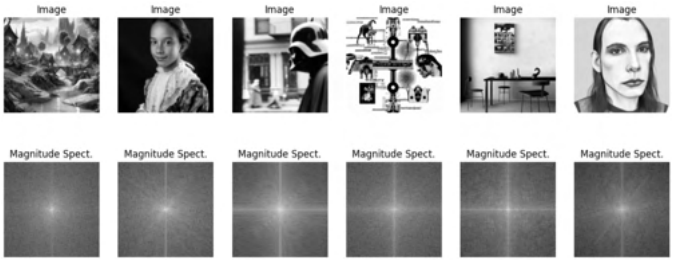

Fig. 5. Sample images and their FFTs from the DALL-E 3 generated content dataset

B. Model Architecture

In order to fairly test the accuracy of the image-only model and the hybrid model, the model architecture for both is consistent with one exception, the number of layers for the input. This is because the image-only model has three input layers, red, green, and blue, while the hybrid model has four

input layers, red, green, blue, and the custom Fast Fourier Transform layer.

The specific model and training architecture for the image-only model is as follows:

- Input: 256×256 pixel images with 3 channels (RGB)
- Conv2D Layer: 32 filters, 3×3 kernel applied on the input shape of 256x256x4
- MaxPooling 2D Layer: Pooling size 2×2
- Conv2D Layer: 64 filters, 3×3 kernel
- MaxPooling 2D Layer: Pooling size 2×2
- Conv2D Layer: 128 filters, 3×3 kernel
- MaxPooling 2D Layer: Pooling size 2×2
- Flatten Layer: Flattens the input and does not affect the batch size
- Dropout Layer: Dropout rate of 0.5 to help prevent overfitting
- Dense Layer: 512 units
- Dense Layer: 1 unit, sigmoid activation function
- Optimizer: Adam optimizer with a learning rate of 0.001
- Number of Epochs: 30
- Batch Size: 20
- Training Steps per Epoch: 100
- Validation Steps per Epoch: 50

The model and training architecture for the hybrid model is as follows:

- Input: 256×256 pixel images with 4 channels (RGB + FFT magnitude spectrum)
- Conv2D Layer: 32 filters, 3×3 kernel applied on the input shape of 256x256x4
- MaxPooling 2D Layer: Pooling size 2×2
- Conv2D Layer: 64 filters, 3×3 kernel
- MaxPooling 2D Layer: Pooling size 2×2
- Conv2D Layer: 128 filters, 3×3 kernel
- MaxPooling 2D Layer: Pooling size 2×2
- Flatten Layer: Flattens the input and does not affect the batch size
- Dropout Layer: Dropout rate of 0.5 to help prevent overfitting
- Dense Layer: 512 units
- Dense Layer: 1 unit, sigmoid activation function
- Optimizer: Adam optimizer with a learning rate of 0.001
- Number of Epochs: 30
- Batch Size: 20
- Training Steps per Epoch: 100
- Validation Steps per Epoch: 50

C. Performance Metric

The metric used to compare model performance is accuracy. The better classification model can be determined by equalizing all reasonable parameters and comparing them with accuracy.

III. Results

The accuracy results for the image-only model can be seen in Figure 6. The accuracy results for the hybrid model can be seen in Figure 7. The final validation accuracy for the image-only model was 72.6%. This can be interpreted to mean that for an image from our dataset to the trained model, the model will accurately classify it 72.6% of the time.

The final validation accuracy for the hybrid model was 84.6%. This can be interpreted to mean that for an image from our dataset to the trained model, the model will accurately classify it 84.6% of the time.

These results indicate that the FFT-integrated hybrid model distinguishes AI-generated content from human-generated content 12.0% more accurately than the image-only model.

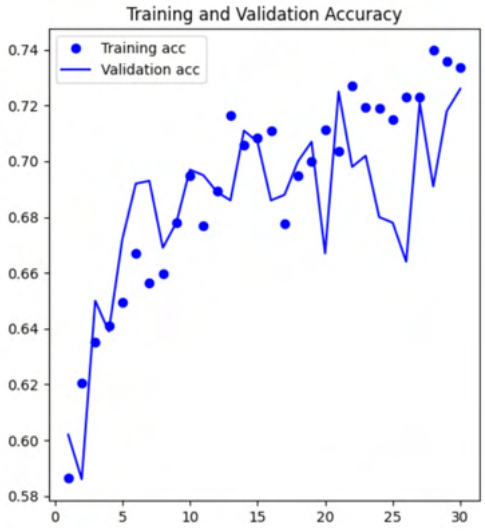

Fig. 6. Image-only model accuracy results

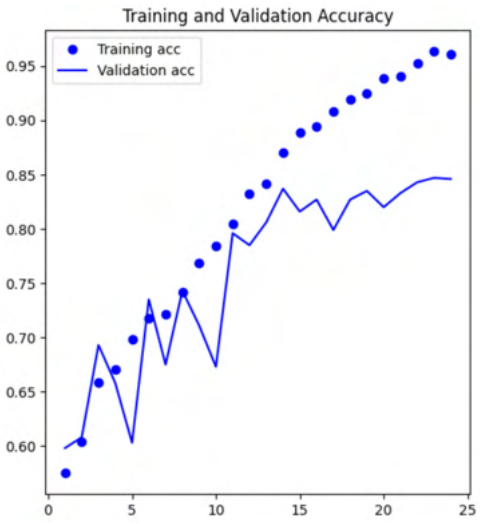

Fig. 7. Hybrid model accuracy results

IV. Discussion

After experimentation, the hybrid model outperforms the image-only model by 12% in accuracy. This indicates that the model is more robust when being trained on both the original image and the FFT of the original image than when just being trained on the image.

The 2D FFT of an image is considered a deterministic function of the original image. It is widely accepted that deterministic functions cannot provide more information to a model than the data it was derived from. The higher accuracy of the hybrid model proves contrary to this widely held belief and was cause for further investigation.

A. Utilizing the Average FFTs

To investigate why the hybrid approach created a more robust model when the extra input layer was a deterministic function of the original images, I computed the average FFT of both classes in hopes of visually identifying a differentiating trend. The average FFT of each class can be seen in Figure 8.

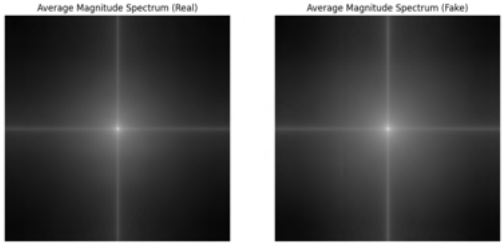

Fig. 8. Average 2D FFT of the human-generated content (left) and DALL-E 3 generated content (right)

Upon observation, the average FFT of the DALL-E 3 generated content dataset has prominent higher frequencies in the FFT that are not as visible in the human-generated content dataset. To observe this phenomenon further, I plotted the difference between the two average FFTs, as seen in Figure 9.

The only areas in which the two dataset's FFTs are similar are where the graph is bright red. This difference in average FFTs indicates that there is a visible distinction between the FFTs of AI-generated content and human-generated content. However, this does not determine the answer as to why a model trained on a deterministic function of the original input in conjunction with the original input performs better.

B. Investigating Differences in FFTs

After observing the difference in average FFTs, especially in the higher frequencies, I wanted to test the hybrid model's dependency on this higher frequency distinction. To do so I applied a Gaussian low-pass filter to 100 images from the DALL-E 3 generated content to feed to the model and observe whether or not the filter could "trick" the model. The Gaussian low-pass filter filters out the higher frequencies of the image before it is inputted into the model. The hybrid model classified only 33 out of the 100 images correctly (33%), compared to the normal validation accuracy of 84.6%. By applying a low-pass filter to the images, I successfully tricked the model into classifying DALL-E-generated images as being human-generated.

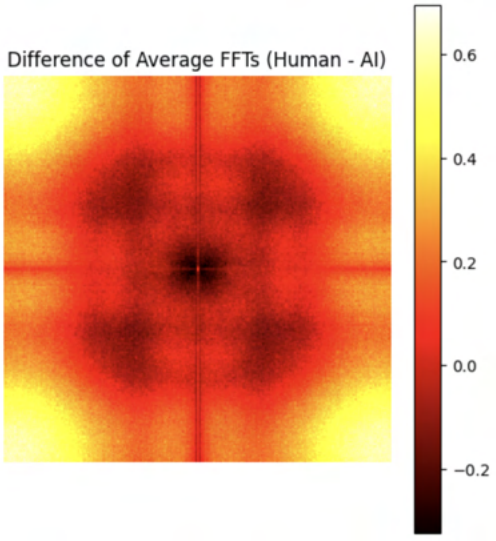

Fig. 9. Difference in Average 2D FFTs

C. Research Implications

Using the magnitude spectra as a means of distinguishing AI images from human images is not novel, as discussed in the introduction. However, the demonstrated increased robustness of a hybrid model, using both the FFT and the original image as input layers is novel and has not been researched or proven before. Given the results of the experimentation, further research is warranted in the applicability of the frequency domain of images in distinguishing AI-created content from other content.

V. Conclusion

Identifying signatures in AI-generated content will be an indispensable capability as AI image-generation models continue to advance. Being able to differentiate between what is AI and what is human will continue to prove more challenging, so as AI capabilities grow, we must grow the techniques we use to combat it.

Further demonstrating trends through the use of the frequency domain in this research presents a new avenue to explore in order for AI-identification techniques to keep up with AI generation capabilities. The novelty of using a hybrid approach with both the original data and the FTT of the original data could be extended further in future research to formats such as video, audio, and beyond.

Further research is also warranted surrounding the ability of the hybrid model to be tricked into misclassification when a Gaussian low-pass filter is applied to the input image.

The generalization of the results must be researched beyond just images generated by DALL-E 3 and from this dataset. Testing this approach with other popular models may prove fruitful if the results hold for other reverse diffusion models

or other mediums. Future research could be done on certain subcategories of images such as images depicting humans or images depicting animals.

Although the original question posed by this research was answered, there is still much more work to be done when it comes to detecting AI-generated content. Through my further research, I hope to explore the applicability of this technique to video and audio as well as to other popular models' images such as Midjourney and Stable Diffusion.

ACKNOWLEDGMENT

I would like to thank my supervisor, Dr. Sanghamitra Dutta, for all of her invaluable help and support throughout my research. Without her, this research would never have come to fruition. I would also like to thank the graduate students on Dr. Dutta's Foundations of Reliable Machine Learning team who provided valuable insights and feedback on my research, specifically Faisal Hamman for his ideas regarding applying a low-pass filter to the DALL-E images.

I would also like to thank Kathryn Weiland in the University of Maryland Electrical and Computer Engineering Department for her support throughout my four years in the program.

REFERENCES

[1] OpenAI, "Dall-e: Creating images from text," https://openai.com/research/dall-e, 2021.

[2] Amazon Web Services, "What is stable diffusion?" https://aws.amazon.com/what-is/stable-diffusion/, 2023.

[3] S. Li, "Diffusion models made easy," Towards Data Science, April 2023, available online: https://towardsdatascience.com/diffusion-models-made-easy-8414298ce4da.

[4] A. Y. J. Ha, J. Passananti, R. Bhaskar, S. Shan, R. Southen, H. Zheng, and B. Y. Zhao, "Organic or diffused: Can we distinguish human art from ai-generated images?" 2024.

[5] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in gan fake images," in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2019, pp. 1–6.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[7] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image generation models," 2023.

[4] [5] [6] [7]