Fourier Feature Attribution: From the View of Signal Decomposition

Anonymous Authors[1]

Abstract

The study of neural networks from the perspective of Fourier features has garnered significant attention. While existing analytical research suggests that neural networks tend to learn low-frequency features, a clear attribution method for identifying the specific learned Fourier features has remained elusive. To bridge this gap, we propose a novel Fourier feature attribution method grounded in signal decomposition theory. Additionally, we analyze the differences between game-theoretic attribution metrics for Fourier and spatial domain features, demonstrating that game-theoretic evaluation metrics are better suited for Fourier-based feature attribution.

Our experiments show that Fourier feature attribution exhibits superior feature selection capabilities compared to spatial domain attribution methods. For instance, in the case of Vision Transformers (ViTs) on the ImageNet dataset, only 8% of the Fourier features are required to maintain the original predictions for 80% of the samples. Furthermore, we compare the specificity of features identified by our method against traditional spatial domain attribution methods. Results reveal that Fourier features exhibit greater intra-class concentration and inter-class distinctiveness, indicating their potential for more efficient classification and explainable AI algorithms.

1. Introduction

Analyzing neural network behavior through the view of Fourier features has proven to be an effective approach(Xu et al., 2019; Xu & Zhou, 2021; Dong et al., 2021; Yin et al., 2019). Numerous studies have demonstrated that neural networks tend to learn low-frequency features(Dong et al., 2021; Xu et al., 2019). Notably, Xu et al(Xu & Zhou, 2021). observed that neural networks initially capture low-frequency features and gradually shift to learning higher-frequency ones. After the invention of Vision Transformer (ViT) models, Dong et al(Dong et al., 2021). further revealed that ViTs exhibit an even stronger preference for low-frequency features. However, these studies fall short of identifying the specific frequency components that neural networks learn.

Motivated by this gap, we propose a novel Fourier feature attribution algorithm based on error response analysis. Unlike feature attribution in the spatial domain, Fourier feature attribution aligns well with game-theoretic evaluation metrics, such as the Deletion-Insertion Game(Srinivas & Fleuret, 2019). Most attribution evaluation metrics are built on the assumption that features can be categorized into those that influence the decision and those that do not(Yang et al., 2023b; Sundararajan et al., 2017; Ancona et al., 2017; Dabkowski & Gal, 2017). If perturbing an influential feature causes a significant change in the network's output, it is deemed important; if the output remains stable when a feature is perturbed, the feature is considered irrelevant. Game-theoretic metrics formalize this idea: if removing a feature does not impact the network's output, it is irrelevant; otherwise, it is significant. However, this assumption faces inherent flaws in the spatial domain(Fel et al., 2023; Srinivas & Fleuret, 2019). From a signal perspective, setting a spatial feature value to zero does not truly "remove" the feature but rather introduces a new signal, which undermines the validity of game-theoretic metrics. This limitation has led to criticisms of game-theoretic attribution evaluation metrics in the spatial domain. In contrast, the Fourier feature space represents signals as a sum of Fourier components. Setting the energy of a specific component to zero naturally aligns with the game-theoretic notion of "removing" a feature, making Fourier feature attribution inherently compatible with these evaluation metrics.

Building on signal theory, we propose a Fourier feature attribution algorithm based on error response analysis. Specifically, we treat the gradient information at the input layer as the error that the input layer needs to satisfy for accurate classification. Leveraging the property that spatial-domain multiplication corresponds to frequency-domain convolution, we quantify the error response of different Fourier features. This allows us to identify the Fourier components

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

that most significantly impact classification.

Our experiments reveal a striking disparity in neural networks' sensitivity to different Fourier features. For unimportant features, the network's output remains virtually unchanged, whereas perturbing important features leads to substantial changes in the output. Notably, only a small subset of Fourier features has a significant influence on the network's predictions. To further compare the specificity of spatial and Fourier feature attributions, we conducted additional experiments to evaluate the inter-class and intra-class variations of the attributed features. The results show that Fourier features exhibit smaller intra-class variations and larger inter-class variations compared to spatial domain features. This indicates that Fourier features are more conducive to effective classification. Our Contributions can be summarized below:

- We evaluate the suitability of game-theoretic evaluation metrics in both Fourier and spatial domain feature spaces, demonstrating that removal-based metrics are better suited for Fourier feature attribution.
- Leveraging signal processing techniques, we propose a Fourier feature attribution method that outperforms most spatial attribution algorithms. Experimental results show that attribution in the Fourier feature space exhibits strong feature selection capabilities.
- Furthermore, our comparative analysis between Fourier and spatial domain feature attribution reveals that Fourier features are better suited for classification tasks due to their superior specificity and distinctiveness.

2. Related Works

2.1. Analyze Neural Network by Fourier Feature

Xu et al.(Xu & Zhou, 2021; Ma et al., 2021), in their theoretical analysis, demonstrated that neural networks first learn low-frequency features before gradually capturing high-frequency ones. Similarly, Bengio(Rahaman et al., 2019). reached a similar conclusion in their experimental studies. With the advent of Vision Transformers (ViTs), Dong further found that ViTs also rely heavily on low-frequency features(Dong et al., 2021). These findings highlight a growing trend of interpreting neural network behavior from the perspective of Fourier features(Yin et al., 2019; Xu et al., 2019), with various methods arriving at similar conclusions. However, unlike the well-established importance evaluation algorithms for spatial features, there has been no comprehensive attribution algorithm to assess the significance of all Fourier features.

2.2. Feature Attribution Algorithms

Existing feature attribution algorithms primarily focus on the spatial domain, aiming to identify pixel-level features that significantly influence a network's decision. These algorithms include gradient-based methods such as Integrated Gradients(Sundararajan et al., 2017), LPI(Yang et al., 2023a), and MIG(Zaher et al., 2024), EG(Erion et al., 2021), perturbation-based methods like SmoothGrad(Smilkov et al., 2017) and LIME(Ribeiro et al., 2016), as well as back-propagation-based methods such as DeepLIFT(Shrikumar et al., 2017), Input×Gradient(Simonyan et al., 2014), guided back-propagation(Springenberg et al., 2015), FullGrad(Srinivas & Fleuret, 2019), and Grad-CAM(Selvaraju et al., 2017). In recent years, hybrid approaches have emerged, combining the strengths of various attribution techniques(Decker et al., 2024). Despite these advancements, the absence of universally accepted evaluation metrics (Yang et al., 2023a; Deng et al., 2024) makes it challenging to determine which attribution method is truly faithful(Dombrowski et al., 2019; Hooker et al., 2019; Zintgraf et al., 2017; Simonyan & Zisserman, 2015). In fact, some methods even produce contradictory explanations for the same model and input, further complicating the evaluation process(Krishna et al., 2024; Neely et al., 2021), indicating that the current attribution methods in the space domain are very unstable(Lin et al., 2023; Zhou et al., 2022; Adebayo et al., 2020; Bilodeau et al., 2022; Fokkema et al., 2023).

Most evaluation metrics are based on the intuitive assumption that features can be divided into those that influence decision-making and those that do not. If perturbing an influential feature causes significant changes in network output, the feature is deemed important; if the output remains stable when a feature is perturbed, the feature is considered non-influential. Methods such as FID(Amara et al., 2022), IR(Rieger & Hansen, 2020), INFD(Yeh et al., 2019), ROAR(Hooker et al., 2019), DIFFID(Yang et al., 2023a), and (Sundararajan et al., 2017) have proposed evaluation metrics grounded in this assumption. However, from a signal composition perspective, many of these metrics inadvertently introduce additional signals when manipulating the original features, undermining their validity. In contrast, game-theoretic evaluation metrics in the Fourier feature space do not introduce extraneous signals. This allows the deletion operation, central to game-theoretic approaches, to remain valid, providing a more robust framework for evaluating feature attribution.

3. Method

3.1. Deletion-Insertion Game

The Deletion and Insertion Game operates on the assumption that removing non-influential signals should lead to

minimal changes in classification confidence while removing important features should cause a significant confidence drop. When removing an equal number of features, the effectiveness of different attribution algorithms is determined by comparing the area enclosed by their confidence variation curves and the coordinate axes.

The algorithm has two versions. In the first version, features deemed unimportant by the attribution algorithm are sequentially removed. The larger the area enclosed by the confidence curve and the coordinate axes, the better the attribution algorithm performs. In the second version, important features are removed sequentially. Here, a smaller enclosed area indicates a more effective attribution algorithm.

However, the use of this metric in the spatial domain presents a fundamental flaw. In works like FullGrad(Srinivas & Fleuret, 2019) and LPI(Yang et al., 2023a), the metric is applied by setting pixel values to zero. From the perspective of signal composition, this operation essentially introduces a pulse signal equivalent to the original pixel value. Removing pixels, therefore, translates into adding pulse signals. Mathematically, this operation can be formulated as:

$$S_{new} = \sum A_i e^{-jw_i} - \sum_{n \in Deletion} \delta_n \tag{1}$$

where δ represents the added pulse at the removed pixel location. Consequently, this operation is not a removing operation but an adding operation in terms of signal processing.

In contrast, within the Fourier feature space, removal can be achieved by setting the corresponding Fourier coefficients to zero, which can be formulated as:

$$S_{new} = \sum A_i e^{-jw_i} - \sum_{n \in Deletion} A_n e^{-jw_n} \tag{2}$$

This operation genuinely represents a removal, distinguishing between the presence and absence of signal components. Therefore, the Deletion and Insertion Game metric, when applied in the Fourier feature space, better aligns with the assumptions of game theory.

3.2. Error Response

This section introduces the details of the algorithm for Fourier feature attribution. We treat the gradient of the network output with respect to the input layer as the error signal of the input layer, meaning that the input data needs to be adjusted in the direction of the back-propagated gradient to align with the network's expected output.

Inspired by the Input×Gradient method for spatial domain attribution, we observe that the Input×Gradient method in the spatial domain corresponds to convolution in the frequency domain. Given that images are discrete signals, we leverage the properties of the discrete Fourier transform (DFT) to obtain the convolution values in the frequency domain:

$$\mathrm{X}_1(\omega_k) * \mathrm{X}_2(\omega_k) = \mathcal{F}(X_1[n] \cdot X_2[n]) \tag{3}$$

where $\mathcal{F}$ denotes DFT, $X[n]$ denotes the signal in the spatial domain, $\mathrm{X}(\omega)$ denotes the signal in the Fourier feature domain. We first construct the error signal by the cross-entropy loss function.

$$\nabla X = \frac{\partial CE(y_i, l_i)}{\partial X} \tag{4}$$

where y_i is the output of the model. Then we obtain the convolution value of each Fourier feature:

$$\mathrm{X}_1(\omega_k) * \mathrm{X}_2(\omega_k) = \mathcal{F}(X[n] \cdot \nabla X[n]) \tag{5}$$

Based on the convolution's properties in signal processing, this new convolution value can be understood as the influence of the error signal on the original signal, forming a new signal that we refer to as the "error response." We then compute the difference between the self-energy of each new signal and that of the original signal as follows:

$$Score = |\mathcal{F}(X[n] \cdot \nabla X[n])| - |\mathcal{F}(X[n])| \tag{6}$$

We use the response strength of the original signal to the error signal to differentiate the importance of various features. Our findings reveal that Fourier components with stronger response intensities are more critical to the network's decision-making process.

4. Experiments

Our experiments aim to address several key questions:

- Q1: Is the attribution provided by our algorithm faithful and accurate?
- Q2: Is the proposed self-energy difference metric effective?
- Q3: What are the differences between the attribution results for Fourier features and spatial features?

To answer these questions, we design a series of experiments.

4.1. Settings

Dataset: We adopt the validation set of the ImageNet2012(Russakovsky et al., 2015) dataset as our dataset. This dataset is also widely used in many feature attribution works.(Kim et al., 2018; Zhou et al., 2018; Srinivas & Fleuret, 2019; Fel et al., 2023)

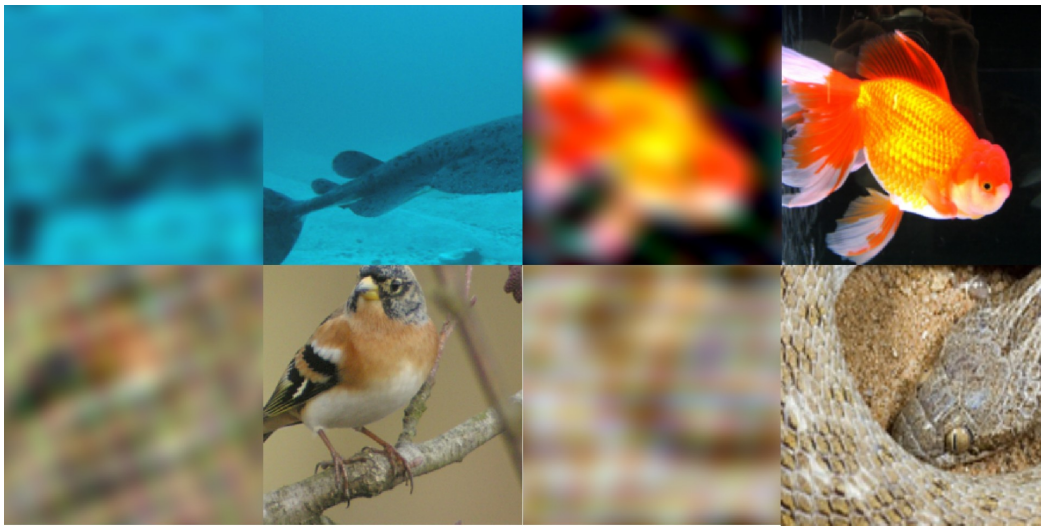
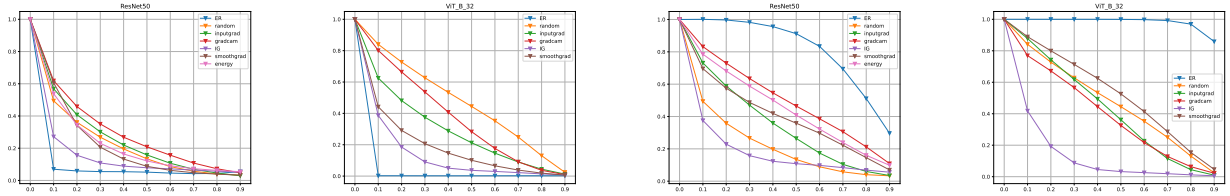


Figure 1. Attribution Visualization: We only preserve 150 Fourier Features in these 4 pictures, however, the ViT_B_32 can still classify them correctly.



(a) ResNet50 Del Most Important (b) ViT Del Most Important (c) ResNet50 Del Least Important (d) ViT Del Least Important

Figure 2. Faithfulness Analysis

Baselines: Integrated-based: IntGrad(Sundararajan et al., 2017), Perturbation-based: SmoothGrad(Smilkov et al., 2017), Gradient×input-based: InputGrad(Simonyan et al., 2014), Other: GradCAM(Selvaraju et al., 2017), FullGrad(Srinivas & Fleuret, 2019). We select the baselines from different types of attribution methods to compare the difference between Fourier Feature Attribution and Space Domain Attribution.

Implementation Details: We implement our experiments on A800 80GB, CentOS8. We implement the IntGrad method by captum(Kokhlikyan et al., 2020). And others are implemented by the publicly available codes.

Backbone Networks: We adopt ResNet50 and ViT_B_32 as our backbone networks. These choices cover both classical and contemporary network architectures, enabling a more comprehensive analysis and comparison of attribution methods' performance.

4.2. Faithfulness Analysis

To address Q1, we design this experiment. Since not all methods are compatible with the architecture of ViT_B_32, FullGrad was excluded from the comparison of this model. Figure2 shows that whether starting from deleting the most important features or the least important features, Fourier feature attribution consistently demonstrated significantly higher faithful performance than spatial feature attribution. Notably, the ViT model retained a high level of confidence even after 90% of the features were removed.

Moreover, the two versions of the deletion-insertion game curves derived from our method better align with the theoretical assumptions of this metric, the curves from both versions should complement each other to form a square. In contrast, certain spatial attribution methods, such as Integrated Gradients, produced nearly identical curves under both metric versions. This strongly suggests that the deletion-insertion game metric is better suited for Fourier feature attribution.

Additionally, from the curves, we observed that the number

of Fourier features supporting ViT decisions is remarkably small. Notably, a significant decline in confidence levels only occurs after deleting 90% of the Fourier features. This finding aligns with previous studies, which have demonstrated that ViT exhibits superior generalization capability compared to ResNet50(Parmar et al., 2019). Therefore, this experiment validates that our attribution method achieves faithful and reliable feature attribution.

4.3. Ablation Study

To address Q2, we design this experiment with the following baseline: transforming the importance scores from spatial attribution methods into the Fourier domain by (inverse)Fourier transform. We then use the magnitude of the transformed value as their scores in the Fourier feature domain. Additionally, We compare the random deletion of Fourier features and the deletion based on energy magnitude.

As shown in Figure3 and 4, when the spatial attribution scores are transformed into the Fourier domain(no matter by FFT or IFFT), the resulting scores become highly unstable, with extreme values oscillating between random deletion and Fourier-guided deletion. This suggests that spatial feature attribution and Fourier feature attribution share a certain degree of correlation. Nevertheless, the performance of these scores still falls short of our proposed method in most cases.

In particular, when comparing against the InputxGrad method, we found that without incorporating the self-energy difference, the Fourier-transformed attribution scores derived from InputxGrad exhibited severe instability, further highlighting the effectiveness of our self-energy difference approach.

By comparing energy-based deletion with random deletion, we observed that although energy-based deletion exhibited lower oscillation, its attribution performance was still suboptimal. This ablation study demonstrates the necessity of introducing both the error response and the self-energy difference. Simply transforming the spatial importance scores into the Fourier domain does show a certain connection to Fourier features but leads to highly unstable faithfulness scores.

4.4. Spatial-Feature and Fourier-Feature

To further address Q3, we conduct experiments focusing on comparing the contribution of attribution results from the two feature spaces to classification tasks. The experiment involved two key evaluations: (1) identifying the minimal number of features required to maintain the network's decision-making and (2) measuring the distinctiveness of features identified by the two attribution methods.

Table 1. ResNet50 Feature Attribution Results

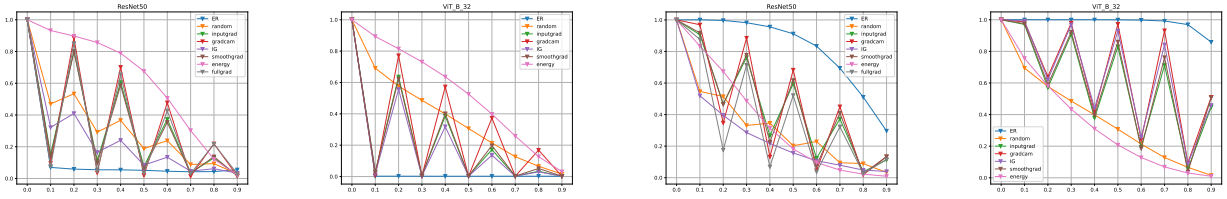
Method	Intra-Class	Inter-Class
ER	**2.935**	**4214**
Fullgrad	1.038	3705
Random	1.0	0
InputxGrad	1.052	1511
GradCAM	1.120	2923
IntGrad	1.0	0
SmoothGrad	1.095	3689

As shown in Figures 2 and 5, Fourier feature selection demonstrated significantly stronger feature filtering capabilities compared to spatial feature attribution. For ViT, only 8% of the Fourier features were sufficient to sustain 80% of sample decisions. Even for ResNet50, which exhibited slightly weaker performance than ViT(Parmar et al., 2019), 10% of Fourier features still maintained decisions for 70% of samples, a feat unattainable with spatial feature attribution.

Feature specificity is analyzed by examining the distribution of high-scoring features within and across classes. High-scoring features were defined as those with scores above the mean value. To eliminate the influence of differing score scales across methods, we normalized high-scoring feature values to 1 and low-scoring features to 0.

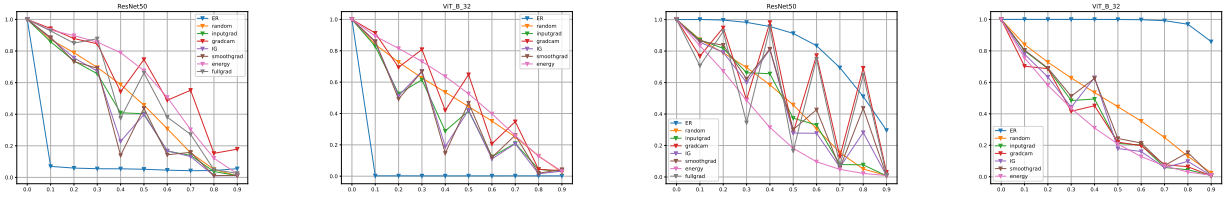
Intra-Class Distribution: This metric reflects the ability to filter noise. The more concentrated the distribution, the stronger the ability to filter noise. Within each class, we aggregated feature scores across all samples along each dimension to compute an overall score. We then calculated the kurtosis of the high-scoring features. The results in Table1 and 2 indicate that Fourier features exhibited a much higher concentration, with kurtosis values approaching 3. In contrast, spatial attribution methods showed much lower concentration, barely exceeding random attribution by 0.2 at best.

Inter-Class Specificity: To assess inter-class specificity, we compute the mean distribution of high-scoring features across different classes. Given the binarization of feature scores to 1 for high and 0 for low, a feature's inter-class mean approaching $1/1000$ indicates greater specificity. We count the number of features with an inter-class mean below $2/1000$ but > 0, as a higher count suggests better specificity. Fourier feature attribution demonstrates a significantly greater number of such features, underscoring its superior ability to capture class-specific distinctions. These two aspects of our experiments demonstrate that Fourier features exhibit both strong intra-class concentration and high inter-class specificity. As shown in Table1 and 2, our method identifies the highest number of features, which leads to superior inter-class specificity in the Fourier feature attribution results. These two experimental aspects demon-



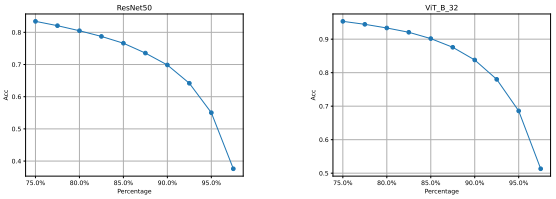
(a) ResNet50 Del Most Important (b) ViT Del Most Important (c) ResNet50 Del Least Important (d) ViT Del Least Important

Figure 3. Ablation Study FFT: The scores are transformed to Fourier feature domain by fast Fourier transformation



(a) ResNet50 Del Most Important (b) ViT Del Most Important (c) ResNet50 Del Least Important (d) ViT Del Least Important

Figure 4. Ablation Study IFFT: The scores are transformed to Fourier feature domain by inverse fast Fourier transformation



(a) ResNet50 (b) ViT_B_32

Figure 5. Max Deletion Rate: In this experiment, we explore the minimal number of features required to maintain the decision-making capability of neural networks.

strate that Fourier features exhibit both strong intra-class concentration and high inter-class specificity. Therefore, Fourier features are better suited for classification tasks and contribute more effectively to the development of explainable AI compared to spatial features.

4.5. Results

Through our experiments, we have addressed the three questions raised at the beginning of this section:
Q1: Under the Deletion-Insertion Game metric, our method demonstrates strong attribution accuracy. It exhibits excellent performance across both versions of the Deletion-Insertion Game for two different models. Moreover, we observe that the Fourier feature attribution curves are better aligned with the theoretical assumptions of the Deletion-Insertion Game, further validating that this metric is particularly well-suited for Fourier feature attribution.
Q2: Based on our ablation study, we confirm the effectiveness of incorporating self-energy differences in our approach.
Q3: Using our proposed inter-class specificity and intra-class concentration metrics, we find that the Fourier feature attribution results exhibit stronger inter-class specificity and higher intra-class concentration. This makes Fourier feature attribution better suited for classification tasks and provides a stronger foundation for building explainable AI systems.

Table 2. ViT Feature Attribution Results

Method	Intra-Class	Inter-Class
ER	**2.934**	**4214**
Random	1.0	0
InputxGrad	1.173	3315
GradCAM	1.032	0
IntGrad	1.0	0
SmoothGrad	1.097	3692

5. Conclusion and Future Work:

This paper proposes a Fourier feature attribution method based on error response, providing a novel perspective for feature attribution through the view of Fourier features. We demonstrate the rationale for applying game-theoretic evaluation metrics in the context of Fourier feature attribution. Ex-

perimental results validate the effectiveness of our method, showcasing the superior feature selection capabilities of Fourier feature attribution. By comparing spatial domain attribution results with those of Fourier feature attribution, we found that Fourier features exhibit better intra-class concentration and inter-class specificity, indicating that Fourier features are more suitable for classification tasks and the development of explainable AI systems.

In the future, we aim to develop explainable AI frameworks based on Fourier feature spaces, enhancing the reliability and robustness of AI systems.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. Debugging tests for model explanations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Amara, K., Ying, Z., Zhang, Z., Han, Z., Zhao, Y., Shan, Y., Brandes, U., Schemm, S., and Zhang, C. Graphframex: Towards systematic evaluation of explainability methods for graph neural networks. In Rieck, B. and Pascanu, R. (eds.), *Learning on Graphs Conference, LoG 2022, 9-12 December 2022, Virtual Event*, volume 198 of *Proceedings of Machine Learning Research*, pp. 44. PMLR, 2022.

Ancona, M., Ceolini, E., Öztireli, A. C., and Gross, M. H. A unified view of gradient-based attribution methods for deep neural networks. *CoRR*, abs/1711.06104, 2017.

Bilodeau, B. L., Jaques, N., Koh, P. W., and Kim, B. Impossibility theorems for feature attribution. *CoRR*, abs/2212.11870, 2022. doi: 10.48550/ARXIV.2212.11870. URL https://doi.org/10.48550/arXiv.2212.11870.

Dabkowski, P. and Gal, Y. Real time image saliency for black box classifiers. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6967–6976, 2017.

Decker, T., Bhattarai, A. R., Gu, J., Tresp, V., and Buettner, F. Provably better explanations with optimized aggregation of feature attributions. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Deng, H., Zou, N., Du, M., Chen, W., Feng, G., Yang, Z., Li, Z., and Zhang, Q. Unifying fourteen post-hoc attribution methods with taylor interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7), 2024. doi: 10.1109/TPAMI.2024.3358410.

Dombrowski, A., Alber, M., Anders, C. J., Ackermann, M., Müller, K., and Kessel, P. Explanations can be manipulated and geometry is to blame. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13567–13578, 2019.

Dong, Y., Cordonnier, J., and Loukas, A. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803. PMLR, 2021.

Erion, G. G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nat. Mach. Intell.*, 3(7), 2021. doi: 10.1038/S42256-021-00343-W.

Fel, T., Picard, A. M., Béthune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., and Serre, T. CRAFT: concept recursive activation factorization for explainability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, 2023. doi: 10.1109/CVPR52729.2023.00266.

Fokkema, H., de Heide, R., and van Erven, T. Attribution-based explanations that provide recourse cannot be robust. *J. Mach. Learn. Res.*, 24:360:1–360:37, 2023.

Hooker, S., Erhan, D., Kindermans, P., and Kim, B. A benchmark for interpretability methods in deep neural networks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 9734–9745, 2019.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, 2018.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., and Reblitz-Richardson, O. Captum: A unified and generic model interpretability library for pytorch. *CoRR*, abs/2009.07896, 2020.

Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., and Lakkaraju, H. The disagreement problem in explainable machine learning: A practitioner's perspective. *Trans. Mach. Learn. Res.*, 2024, 2024.

Lin, C., Covert, I., and Lee, S. On the robustness of removal-based feature attributions. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Ma, Y., Xu, Z. J., and Zhang, J. Frequency principle in deep learning beyond gradient-descent-based training. *CoRR*, abs/2101.00747, 2021.

Neely, M., Schouten, S. F., Bleeker, M. J. R., and Lucic, A. Order in the court: Explainable AI methods prone to disagreement. *CoRR*, abs/2105.03287, 2021.

Parmar, N., Ramachandran, P., Vaswani, A., Bello, I., Levskaya, A., and Shlens, J. Stand-alone self-attention in vision models. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 68–80, 2019.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F. A., Bengio, Y., and Courville, A. C. On the spectral bias of neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 2019.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should I trust you?": Explaining the predictions of any classifier. In Krishnapuram, B., Shah, M., Smola, A. J., Aggarwal, C. C., Shen, D., and Rastogi, R. (eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778.

Rieger, L. and Hansen, L. K. IROF: a low resource evaluation metric for explanation methods. *CoRR*, abs/2003.08747, 2020.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3), 2015. doi: 10.1007/S11263-015-0816-Y.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017*, 2017. doi: 10.1109/ICCV.2017.74.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, 2017.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. URL http://arxiv.org/abs/1706.03825.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. A. Striving for simplicity: The all convolutional net. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

Srinivas, S. and Fleuret, F. Full-gradient representation for neural network visualization. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on*

Neural Information Processing Systems 2019, NeurIPS 2019, 2019.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, 2017.

Xu, Z. J. and Zhou, H. Deep frequency principle towards understanding why deeper learning is faster. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 10541–10550. AAAI Press, 2021. doi: 10.1609/AAAI.V35I12.17261.

Xu, Z. J., Zhang, Y., and Xiao, Y. Training behavior of deep neural network in frequency domain. In Gedeon, T., Wong, K. W., and Lee, M. (eds.), *Neural Information Processing - 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12-15, 2019, Proceedings, Part I*, volume 11953 of *Lecture Notes in Computer Science*, pp. 264–274. Springer, 2019. doi: 10.1007/978-3-030-36708-4_22.

Yang, P., Akhtar, N., Wen, Z., and Mian, A. Local path integration for attribution. In Williams, B., Chen, Y., and Neville, J. (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence*, 2023a. doi: 10.1609/AAAI.V37I3.25422.

Yang, P., Akhtar, N., Wen, Z., Shah, M., and Mian, A. S. Re-calibrating feature attributions for model interpretation. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, 2023b.

Yeh, C., Hsieh, C., Suggala, A. S., Inouye, D. I., and Ravikumar, P. On the (in)fidelity and sensitivity of explanations. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 10965–10976, 2019.

Yin, D., Lopes, R. G., Shlens, J., Cubuk, E. D., and Gilmer, J. A fourier perspective on model robustness in computer vision. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13255–13265, 2019.

Zaher, E., Trzaskowski, M., Nguyen, Q., and Roosta, F. Manifold integrated gradients: Riemannian geometry for feature attribution. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Zhou, B., Sun, Y., Bau, D., and Torralba, A. Interpretable basis decomposition for visual explanation. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), *Computer Vision - ECCV 2018 - 15th European Conference*, volume 11212 of *Lecture Notes in Computer Science*, 2018. doi: 10.1007/978-3-030-01237-3_8.

Zhou, Y., Booth, S., Ribeiro, M. T., and Shah, J. Do feature attribution methods correctly attribute features? In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 9623–9633. AAAI Press, 2022. doi: 10.1609/AAAI.V36I9.21196.

Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. Visualizing deep neural network decisions: Prediction difference analysis. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.