The Role of Frequency in Shaping Features from Artificial Vision Models

Thomas Garity[⋆1] **Thomas Fel**[⋆2] **George A. Alvarez**[1] **Thomas Serre**[2]
[1] Harvard University [2] Brown University
[⋆] Authors contributed equally to the work

Abstract

In our study, we analyzed over 150 state-of-the-art vision models using explainability tools to see how they process high- and low-frequency features. We introduce a metric based on Attribution methods to quantify the models' dependence on high-frequency features. We found that more advanced models rely more on low-frequency features. To advance our investigation, we assessed whether more accurate models demonstrate increased reliance on phase information, which is crucial for human recognition. This was achieved by mixing the phase components of images to evaluate the models' object recognition capabilities. The findings indicate that while high-performing models are progressively depending more on phase information, they substantially lag behind human performance. Additionally, we show that models that depend on low-frequency features tend to have a shape bias, confirming a connection between frequency reliance and perception bias.

Our analysis indicates that as models become more performant, their use of phase information and low-frequency features increases. However, a significant gap remains compared to human capabilities, suggesting opportunities for further enhancing model alignment through frequency analysis.

Keywords: Explainability, Fourier Analysis, Image Recognition, Deep Learning, Vision Models

This work seeks to better understand the features learned by Deep Neural Networks from a frequency perspective. We specifically aim to address three issues: (***i***) the predominance of high-frequency versus low-frequency features in DNNs, (***ii***) the extent of phase information utilization in comparison to human vision, and (***iii***) the influence of frequency analysis to understand the shape-texture bias.

High Frequency Reliance We first ask: do models use high or low-frequency features to drive decisions? To answer this, we will use attribution methods, more specifically, Saliency [16], Integrated Gradients [13], and SmoothGrad [11]. These methods take given image $\boldsymbol{x} \in \mathbb{R}^{w\times h}$ and its prediction $\boldsymbol{f}(\boldsymbol{x})$ as inputs and return a heatmap revealing the pixels considered important for the model, i.e., the areas on which the important features rely. We study these important features from a frequency point of view, to know on average if the model uses low or high-frequency features. More formally, an attribution method [16, 2, 1, 11, 13, 10] returns a heatmap $\gamma \in \mathbb{R}^{w\times h}$. We respectively denote $\mathcal{F}$ and $\mathcal{F}^{-1}$ the 2-D Discrete Fourier Transform (DFT) on some image $\boldsymbol{x}$ and its inverse, such that $\boldsymbol{x} = \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{x}))$. With

Figure 1: **Phase Switching Experiment:** A visual representation of the methodology used to assess the impact of phase information on model performance. This figure illustrates the process of interchanging phase components between images to evaluate the model's reliance on phase information.

$\boldsymbol{z} = \mathcal{F}(\boldsymbol{x}) \in \mathbb{C}^{w\times h}$, the centered Fourier spectrum (freq 0 at the center), we denote the polar form of $\boldsymbol{z} = \boldsymbol{r}e^{i\boldsymbol{\varphi}}$. Finally, we denote $\mathcal{F}(\boldsymbol{x})(i,j) \in \mathbb{C}$ the component i,j of the matrix $\boldsymbol{z}$ of the Fourier spectrum. The metric that we propose consists of calculating the average of the Fourier spectrum of the heatmaps given by an attribution method, to have, on average, the type of frequencies that the model uses. It then remains to calculate the energy allocated by the model in high and low frequencies. We recall that we obtain the energy for a frequency band ω named $E(\omega)$ for the frequency ω via *azimuthal integration*:

$$E(\omega;\gamma) = \int_0^{2\pi} ||\mathcal{F}(\gamma)(\omega\cdot\cos(\theta),\omega\cdot\sin(\theta))||^2 \,\mathrm{d}\theta \quad (1)$$

$$\text{with } \omega \in \{1,\dots,h/2-1\} \quad (2)$$

Subsequently, we introduce our normalized energy metric:

$$\overline{E}(\omega;\gamma) = \frac{E(\omega;\gamma)}{\sum_{\omega'} E(\omega';\gamma)} \quad (3)$$

This enables us to present our metric for measuring high frequencies, essentially the energy of a frequency weighted by the frequency value itself, across the normalized spectrum.

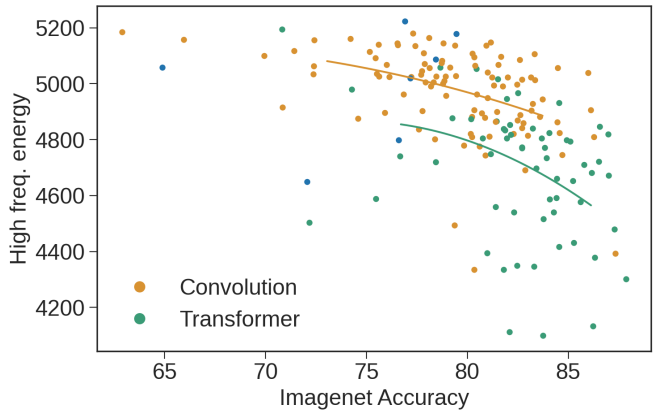

Figure 2: **High-Frequency Energy (avg. $\Lambda(\gamma)$ score) vs. ImageNet Accuracy:** This graph showcase the inverse relationship between model accuracy and reliance on high-frequency features. Notably, transformer models seems to demonstrate a pronounced preference for low-frequency features over convolutional networks.

This approach ensures that signals with higher energy are not disproportionately penalized. This score, on average, indicates the rate of high-frequency features utilized by a given model $\boldsymbol{f}$:

$$\Lambda(\gamma) = \sum_{\omega} \overline{E}(\omega;\gamma) \cdot \omega \tag{4}$$

We employ the Timm library [15] to assess over 150 state-of-the-art models. The outcomes, as illustrated in Figure 2, indicate that the more accurate on ImageNet the models are, the more they depend on low-frequency features. Interestingly, a slight trend towards superior performance is observed in transformer models, which, at equal accuracy levels, appear to utilize fewer high-frequency components.

We conclude that performant models use the more central portion of the Fourier spectrum. However, we still must ask: do they, like humans, rely on low-frequency phase information [8, 14, 9], or do they leverage another type of information?

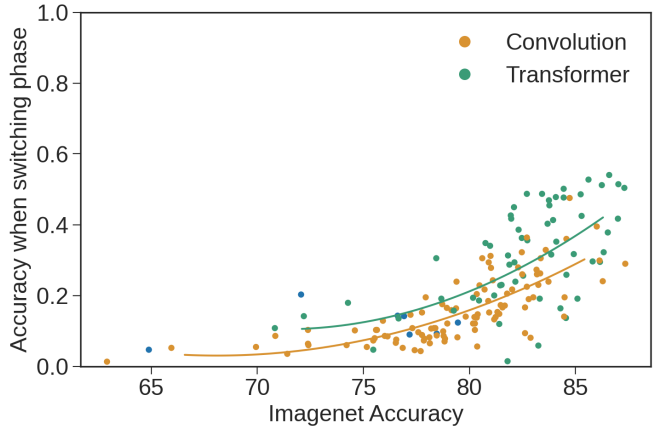

Figure 3: **Phase mixing results.** The more performant models are, the more they rely on phase information. It is also observed that transformer-type models seem to rely more on phase in general compared to convnets.

Phase mixup experiment. We now have a tendency: more performant models increasingly use low frequencies. Moreover, Oppenheim & Lim [8] showed that the Fourier phase spectrum is more important to perception of natural images than the magnitude data, but can we say the same about artificial models? We set up a test to verify if models that use lower frequencies also use more phase information.

In our study, we applied the same models to a subset of 5000 images from the ImageNet test set for which the phase was mixed, as seen in 1. Subsequently, we evaluated the accuracy of the models ($\mathbb{P}_{\boldsymbol{x}_1,\boldsymbol{x}_2'}(\boldsymbol{f}(\boldsymbol{x}_1) = \boldsymbol{f}(\boldsymbol{x}_2')$, in Fig.3) on these modified images, with the results also presented in Fig.3. Our findings reveal two key insights: (1) as a model's accuracy enhances, it increasingly relies on low-frequency features and (2) there's a notable shift towards greater utilization of phase information. This indicates that as models get better, they begin to focus on low-frequency phases, aligning more closely with some results in human perception [8].

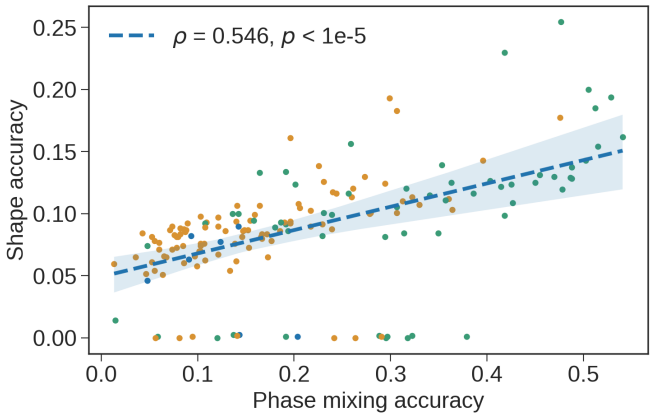

Figure 4: **Correlation between phase mixing score and shape bias on [4].**

Connection with Shape vs. Texture Bias. Current studies reveal a preference for texture over shape in AI models [4, 3, 5, 6, 7] and a link between frequency analysis and shape bias [12]. Here we investigate if shape-biased models also lean on lower frequency features. We examined 150 models using the Cue-Conflict dataset [4], where images blend shapes and textures from different classes. Results presented in Figure 4 highlight a significant relationship between the shape bias of models and their performance in our phase-mixing experiment. This suggests that analyzing models based on frequency can provide a fresh viewpoint on the texture bias of modern DNN.

Conclusion. Our study reveals that while accurate models lean towards low-frequency features and are using more phase information, they still lag significantly behind humans in simple phase-switching tasks – achieving only about 50% of human performance at best. This underscores substantial potential for improvement in vision model development. Employing frequency analysis could offer valuable insights into ongoing question like shape versus texture.

References

Fel, T., Cadene, R., Chalvidal, M., Cord, M., Vigouroux, D., & Serre, T. (2021). Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In *Advances in neural information processing systems (neurips).*

Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the ieee international conference on computer vision (iccv).*

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hermann, K., Chen, T., & Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*.

Hermann, K., & Lampinen, A. (2020). What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems (NeurIPS)*.

Hermann, K. L., Mobahi, H., Fel, T., & Mozer, M. C. (2023). On the foundations of shortcut learning. In *International conference on learning representations.*

Oppenheim, A. V., & Lim, J. S. (1981). The importance of phase in signals. *Proceedings of the IEEE*, *69*(5), 529–541.

Piotrowski, L. N., & Campbell, F. W. (1982). A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K.-R. (2016). Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. In *Workshop on visualization for deep learning, proceedings of the international conference on machine learning (icml).*

Subramanian, A., Sizikova, E., Majaj, N., & Pelli, D. (2024). Spatial-frequency channels, shape bias, and adversarial robustness. *Advances in Neural Information Processing Systems (NeurIPS)*.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the international conference on machine learning (icml).*

Thomson, M. G., Foster, D. H., & Summers, R. J. (2000). Human sensitivity to phase perturbations in natural images: a statistical framework. *Perception*.

Wightman, R., et al. (2019). *Pytorch image models.*

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of the ieee european conference on computer vision (eccv).*