I-ASIDE: Towards the Global Interpretability of Image Model Robustness through the Lens of Axiomatic Spectral Importance Decomposition

Anonymous authors
Paper under double-blind review

Abstract

Robust decisions leverage a high proportion of robust features. Natural images have spectral anisotropy and the majority of spectral energy concentrates on low-frequency components. A change with an infinitesimal amount of energy on the high-frequency components can rewrite the features dominated by high-frequency components. Image models are parameterized general non-linear signal filters. The fragility of the learned feature representations of image models correlates with spectral structures. The spectral importance decomposition of the statistical expectations of the negative decision risks of models with respect to spectrum can thus reflect model robustness from the perspective of feature robustness. To this end, we formulate the spectral importance decomposition problem, and, present **I**mage **A**xiomatic **S**pectral **I**mportance **D**ecomposition **E**xplanation (**I-ASIDE**) – a model-agnostic global interpretability method – to quantify model global robustness from the perspective of the susceptibility of feature representations to perturbations. Our approach provides a unique insight into interpreting model global robustness and enables a considerable number of applications in research, from measuring model robustness, to studying learning dynamics, to assessing label noise, to investigating adversarial vulnerability, etc. We also showcase multiple applications across multiple research domains to endorse such claims.

1 Introduction

Global interpretability (Lipton, 2018) summarizes the decision dynamics of neural networks *en masse* in contrast to instance-wise local interpretability. Local interpretability for image models has achieved great success (Sundararajan et al., 2017; Smilkov et al., 2017; Linardatos et al., 2020; Selvaraju et al., 2017; Arrieta et al., 2020; Zhou et al., 2016; Ribeiro et al., 2016; Lundberg & Lee, 2017; Lakkaraju et al., 2019; Guidotti et al., 2018; Bach et al., 2015; Montavon et al., 2019; Shrikumar et al., 2017), yet quantifiable global interpretability remains virtually unexplored. A brief literature review regarding global interpretability for image models is provided in Section 2.

The global robustness of models reflects an intrinsic property of models and delineates a crucial aspect towards interpretability and trustworthiness. To this end, we present **I-ASIDE**[1][2], a model-agnostic method, to quantify the global robustness of image models from the perspective of the interactions of robust features and non-robust features in decisions from the perspective of spectral importance. Unlike prior works, **I-ASIDE** directly quantifies model global robustness and enables a considerable number of applications in deep learning research across multiple domains (See Section 4).

Feature representations are the maps associated with supervision signals learned by models in the training process. The robustness of learned feature representations largely delimits decision robustness afterwards. Robust decisions leverage more robust features while non-robust decisions leverage less robust features. The fragility of learned features correlates with the energy distributions of spectral structures from the perspective

[1]Anonymized reproducibility: `https://anonymous.4open.science/r/IASIDE_paper_reproducibility-BB9F/`.
[2]The full showcase experiments have consumed 18000+ core hours on supercomputing center with multiple GPUs to complete.

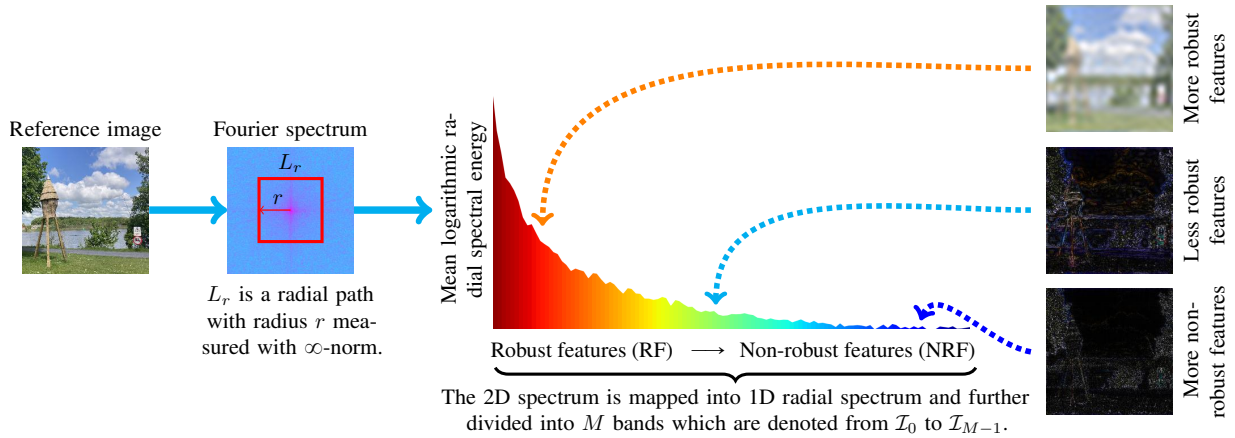


Figure 1: This diagram shows the relationship between the anisotropic spectral structures of natural images and the robustness of feature representations. According to Parseval identity theorem, small spatial perturbations can rewrite the feature representations dominated by high-frequency components.

of the susceptibility of feature representations to perturbations. Natural images have spectral anisotropy and the majority of spectral energy concentrates on low-frequency components (See Figure 1). According to the Parseval identity theorem – the energy of the perturbations in the spatial domain and the spectral domain is conservative and equivalent; small spatial perturbations can rewrite those feature representations dominated by high-frequency components. Consequently, the signal spectrum can index the robustness of feature representations.

Ilyas et al. use the terms 'robust feature (RF)' and 'non-robust feature (NRF)' to theoretically analyze the adversarial robustness problem from the perspective of feature representations and argue that the presence of non-robust features can incur model robustness issues (Ilyas et al., 2019; Tsipras et al., 2018). We adopt the use of their terms and step forward to rigorously discuss the robust decision problem in a broader scope beyond adversarial robustness.

Measuring the ratio of the robust features in decisions can help to interpret model inference dynamics globally from the perspective of feature representations. Unfortunately, discriminating features between being robust and being non-robust is difficult. This research is inspired by the power-law like spectral structures of natural images and feature robustness can be indexed by radial spectrum (See Figure 1). We refer to the radial spectrum as the 'robustness spectrum' thereafter. The decompositions of the statistical expectations of the negative decision risks[3] of models with respect to robustness spectrum can summarize model decision robustness, and quantitatively reflect how neural networks respond to signals in the frequency-domain.

The spectral decompositions are a unique solution satisfying a set of desirable axioms. We formulate this problem by taking four fair distribution axioms from Shapley value theory (Roth, 1988; Aumann & Shapley, 2015) and adapting them into a strong version: *strong efficiency*, *symmetry*, *linearity*, and the *null player* (See Section 3.2). Unlike the ordinary representation of Shapley value theory, the spectral decomposition is represented as a linear equation system in terms of the statistical expectation of some decision risk function. **I-ASIDE** is implemented by solving the spectral importance decomposition equation (7). Our approach provides a unique insight into interpreting the global decision robustness of image models and enables multiple applications in research such as:

- Quantifying model global robustness by examining spectral importance distributions (See the experiments regarding spectral importance distributions in Figure 3, summarized numerical comparison and t-SNE projection in Figure 4);

[3]Larger decision risk values have lower prediction confidences for fixed labels.

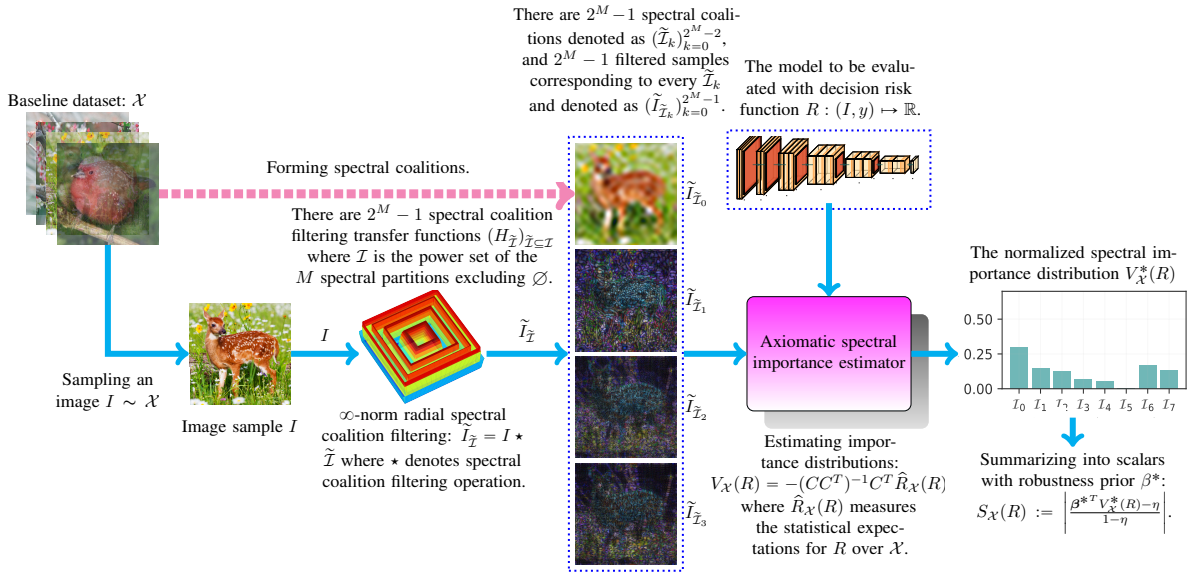


Figure 2: This diagram illustrates the overview of **I-ASIDE** with some baseline dataset $\mathcal{X}$ and some model with the decision risk function $R:(I,y)\mapsto\mathbb{R}$. Image samples are sampled from $\mathcal{X}$. The sampled images are filtered with the spectral coalition filtering consisting of 2^M-1 filtering transfer functions to derive the 2^M-1 filtered samples exclusively containing the specific features indicated by $\widetilde{\mathcal{I}}\subseteq\mathcal{I}$. The filtered samples are used to compute the spectral contributions by using the **A**xiomatic **S**pectral **I**mportance **D**ecomposition Equation 7.

- Understanding the learning behaviours of image models on the datasets with supervision noise (See the experiments in Figure 5);
- Investigating the learning dynamics of models from the perspective of the evolution of feature representation robustness in optimization (See the experiments in Figure 6);
- Providing an insight into understanding the adversarial vulnerability of models by examining spectral importance distributions (See the experiments in Figure 7 and the full results are provided in supplementary material).

It is also notable that the potential applications are not restricted to the above fields. Other research fields such as data augmentation and self-supervised learning can also use **I-ASIDE** as a device to understand the dynamics of the learned feature representations of models. For example, in data augmentation research, **I-ASIDE** can interpret the training dynamics with ablation experiments by understanding how hyper parameters and augmentation tricks can affect the robustness of the learned features.

2 Related work

We summarize related works from three categories to provide research context in visual models: (1) Global interpretability, (2) model robustness and (3) frequency-domain research for learning dynamics.

Global interpretability: Global interpretability summarizes the decision behaviours of models and provides a holistic view. Feature visualization with neuron or class activation maximization can show the ideal inputs for specific neurons or classes by optimizing inputs, and help to understand the learned feature representations of models (Olah et al., 2017; Nguyen et al., 2019; Zeiler et al., 2010; Simonyan et al., 2013; Nguyen et al., 2016a;b). Yet, neural networks are surjective maps and the results often yield 'surrealistic' inputs which are not interpretable. Network dissection attempts to establish the connection between the functions of the units (e.g. a set of channels or a set of layers) in convolutional neural networks and some concepts – e.g. eyes or ears (Bau et al., 2017). Concept-based approach measures the activations of networks with respect to

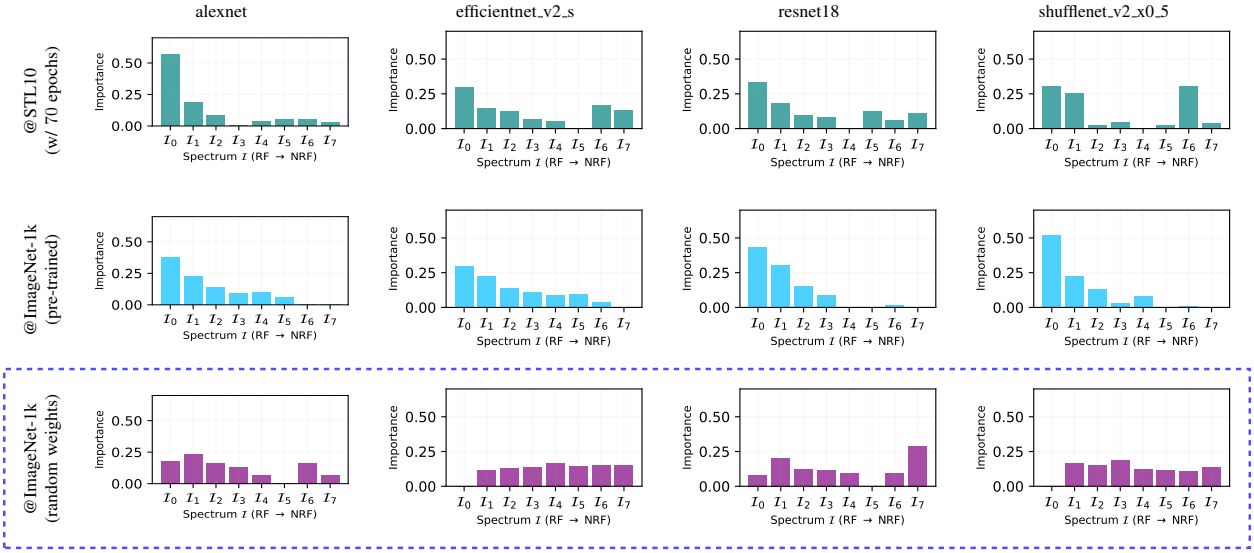


Figure 3: We showcase the spectral importance distributions from multiple models trained on *STL10* with 70 epochs and pre-trained on *ImageNet-1k* respectively. We also include the models with random weights as a control marked by blue box. There are three findings: (1) The spectral importance of the robustness of feature representations of trained models is anisotropic, (2) the spectral importance of the robustness of the feature representations of un-trained models is isotropic, and (3) the models trained on larger datasets (e.g. the middle row) learn more robust feature representations.

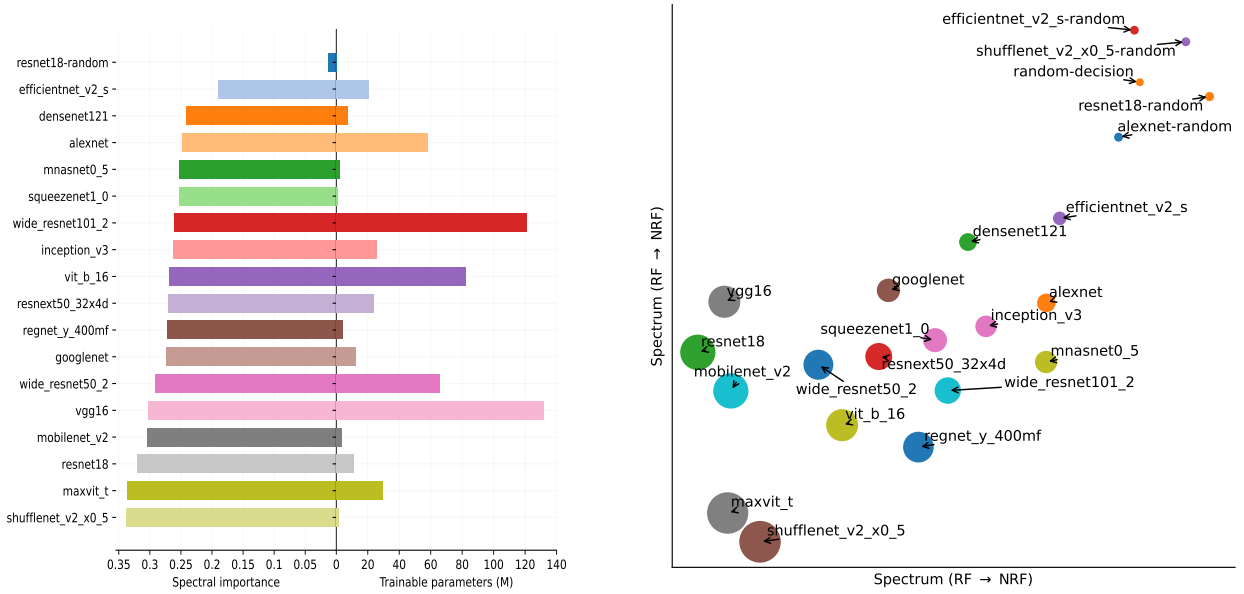
concepts, and interpret networks at concept level (Kim et al., 2018; Ghorbani et al., 2019; Koh et al., 2020; Chen et al., 2020).

Model robustness: Szegedy et al. notice the robustness problem of deep models arising from adversarial perturbations. Thereafter, empirical observations demonstrate that boosting model robustness with adversarial training is at the cost of model standard accuracy for visual tasks (Goodfellow et al., 2014). Later theoretical analyses show standard accuracy is at odds with model robustness (Zhang et al., 2019; Tsipras et al., 2018). Ilyas et al. argue features can be distinguished by their brittleness from robust features and non-robust features, and show that adversarial robustness relates to non-robust features. Our research is based on the above insights.

Frequency-domain research: Neural networks are non-linear parameterized signal processing filters. Investigating how neural networks respond to inputs in the the frequency-domain can provide a unique insight into understanding its functions. Rahaman et al. approximate ReLU networks with piece-wise continuous linear functions in order to perform Fourier analysis. Their results suggest neural networks have a 'spectral bias' on smooth hypotheses (Raghu et al., 2017; Montufar et al., 2014; Rahaman et al., 2019). Xu et al. investigate the learning dynamics of neural networks on frequency-domain in their work 'F-Principle' (Xu et al., 2019a;b). Their work suggests that the learning behaviors of neural networks are spectral anisotropic: Neural networks fit low-frequency components first, then high-frequency components later. Tsuzuku & Sato affirm convolutional neural networks have spectral anisotropy regarding Fourier bases (Tsuzuku & Sato, 2019). Wang et al. conducts an empirical study of the connection between supervision signals and feature representations from the perspective of frequency (Wang et al., 2020). Their works motivate this work to treat neural networks as signal processing filters and investigate how neural networks respond to signals in the frequency-domain.

3 Axiomatic spectral importance decomposition

Figure 2 shows the overview of **I-ASIDE**. We derive the decomposition equation from a set of stronger desirable fair distribution axioms adapted from the ordinary axioms in Shapley value theory (Roth, 1988).



(a) Spectral importance score comparison

(b) Spectral importance distribution projection

Figure 4: This is an application showcase (A1) in research with two experiments. The left bidirectional chart showcases to model spectral importance scores with respect to trainable parameters. The right figure showcases to visualize model robustness by projecting the spectral importance distributions with t-SNE. The shape sizes correspond to robustness (the larger is the better). The experiments are performed on dataset *ImageNet*. The results also correlate with the experiments in Figure 7 regarding adversarial perturbations. In our experiments, the numbers of the trainable parameters of models do not play a crucial role on model robustness.

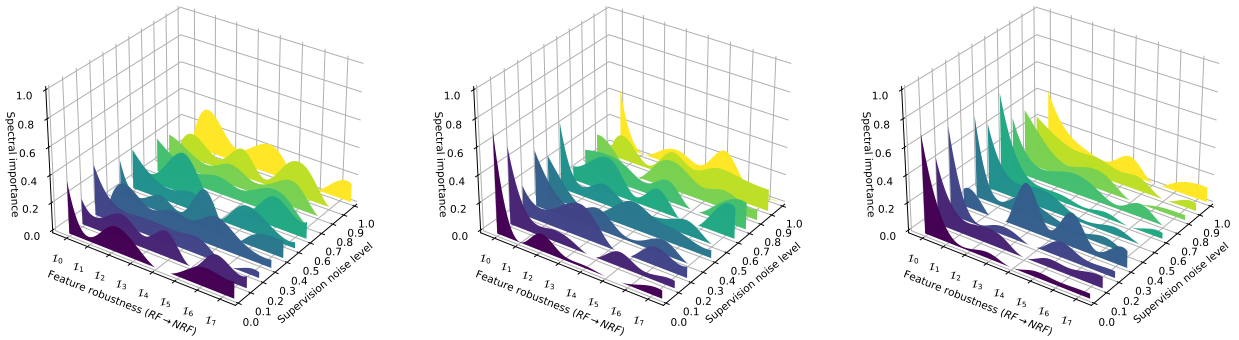
3.1 Setting

We use the coalitional game theory language. Let $[0, 1]$ be the normalized 1-dimensional radial spectrum. The radial spectrum is partitioned into M equispaced regions in which each 1-dimensional radial spectral region corresponds to a 2-dimensional spectral region in a 2-dimensional image spectrum (See Figure 1). Every partition is a '*spectral player*' and the i-th spectral player is denoted as $\mathcal{I}_i$. The M spectral players consist of a player set (*partially ordered set*) indexed by feature robustness and such a player set is denoted as $\mathcal{I} = \{\mathcal{I}_i\}_{i=0}^{M-1}$ (where $M = |\mathcal{I}|$). Spectral players *cooperate*, *interact*, and *contribute* to the decisions of neural networks with various bargaining powers. A '*spectral player coalition*' $\widetilde{\mathcal{I}}$ is a subset of $\mathcal{I}$. A gain measure function $\mu : \widetilde{\mathcal{I}} \mapsto \mathbb{R}$ measures the gain for some coalition $\widetilde{\mathcal{I}}$. The contributions of some coalition $\widetilde{\mathcal{I}}$ with $|\widetilde{\mathcal{I}}|$ players regarding some gain measure function μ are denoted as $V_{\widetilde{\mathcal{I}}}(\mu) := \left(V_{\widetilde{\mathcal{I}},i}(\mu)\right)_{i=0}^{|\widetilde{\mathcal{I}}|-1}$[4]. In this paper, the gain functions are the statistical expectation of the negative decision risk function of some neural network over some baseline dataset $\mathcal{X}$.

3.2 Decomposition axioms

The fair decompositions of the statistical expectations of the negative decision risks can reflect the spectral importance in decisions and should satisfy a set of desirable axioms: *strong efficiency*, *symmetry*, *linearity* and *null player* (Roth, 1988; Hart, 1989; Winter, 2002).

[4]The notation $\boldsymbol{a} := (a_i)_{i=0}^{d-1}$ denotes a d-dimensional vector $\boldsymbol{a}$ in which the i-th member is a_i.



(a) *googlenet* (b) *resnet18* (c) *vit*

Figure 5: This is an application showcase (A2) demonstrating how models respond to label noise in training. We create the noisy-label datasets from clean *Caltech101* by randomly re-assigning a proportion of labels. We vary the proportions (label noise levels) from 0 to 1 with stride 0.1. We train *googlenet*, *resnet18* and *vit* on the derived noisy-label datasets with 200 epochs, learning rate 0.0025 and SGD optimizer. The image sizes are set to 64×64, 64×64 and 224×224 respectively. There are two findings: (1) The spectral importance of the learned feature representations of models is more anisotropic with low supervision signal noise levels and (2) the spectral importance gradually loses anisotropy and exhibits more isotropic.

***Strong efficiency* axiom**[5]: This axiom states that spectral importance decompositions must be σ-additive (Halmos, 2013) on set $\mathcal{I}$. Let $\left(V_{\widetilde{\mathcal{I}},i}(\mu)\right)_{i=0}^{|\widetilde{\mathcal{I}}|-1}$ be the contributions of the $|\widetilde{\mathcal{I}}|$ players on coalition $\widetilde{\mathcal{I}}$. The axiom states that the contributions $\left(V_{\widetilde{\mathcal{I}},i}(\mu)\right)_{i=0}^{|\widetilde{\mathcal{I}}|-1}$ must be summed to $\mu(\widetilde{\mathcal{I}})$ such that: $\sum_{i=0}^{|\widetilde{\mathcal{I}}|-1} V_{\widetilde{\mathcal{I}},i}(\mu) = \mu(\widetilde{\mathcal{I}})$.

***Symmetry* axiom**: Let $\widetilde{\mathcal{I}}$ be some spectral player coalition. Let $\mathcal{I}_a, \mathcal{I}_b \notin \widetilde{\mathcal{I}}$ be two spectral players. The condition $\mu(\widetilde{\mathcal{I}} \cup \{\mathcal{I}_a\}) = \mu(\widetilde{\mathcal{I}} \cup \{\mathcal{I}_b\})$ implies that the decomposed contributions for $\mathcal{I}_a$ and $\mathcal{I}_b$ must satisfy the '*equal treatment of equals*' principle such that: $V_{\{\mathcal{I}_a\}}(\mu) = V_{\{\mathcal{I}_b\}}(\mu)$.

***Linearity* axiom**: Let μ and ρ be two gain measure functions. Let $\widetilde{\mathcal{I}}$ be some spectral player coalition. The axiom states that the decomposition must satisfy $V_{\widetilde{\mathcal{I}}}(\mu + \rho) = V_{\widetilde{\mathcal{I}}}(\mu) + V_{\widetilde{\mathcal{I}}}(\rho)$ and $V_{\widetilde{\mathcal{I}}}(k\mu) = kV_{\widetilde{\mathcal{I}}}(\mu)$ where $(\mu + \rho)(\widetilde{\mathcal{I}}) := \mu(\widetilde{\mathcal{I}}) + \rho(\widetilde{\mathcal{I}})$ and $k \in \mathbb{R}$.

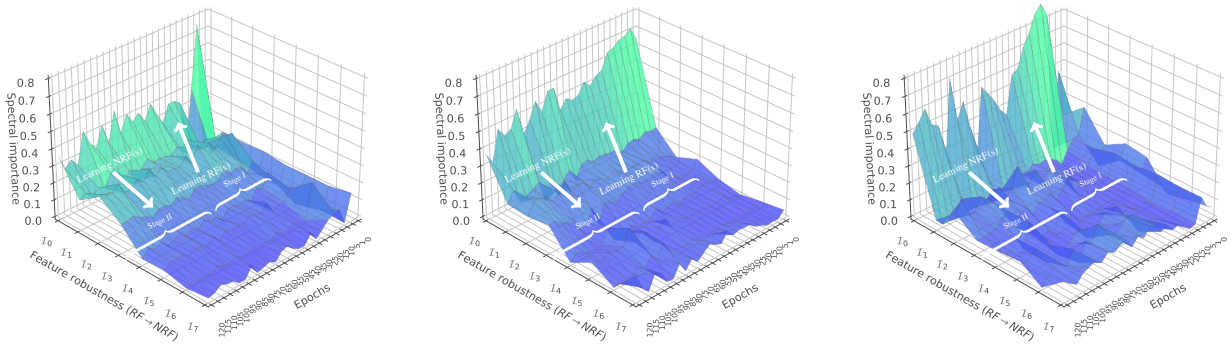
***Null player* axiom**: A null player $\mathcal{I}^*$ is the player who has no contribution such that: $V_{\{\mathcal{I}^*\}}(\mu) = 0$ and $V_{\widetilde{\mathcal{I}} \cup \{\mathcal{I}^*\}}(\mu) = V_{\widetilde{\mathcal{I}}}(\mu)$ for $\forall \mathcal{I}^* \notin \widetilde{\mathcal{I}} \subseteq \mathcal{I}$.

3.3 Decision risk function design

We use the statistical expectations of negative decision risks to measure the 'gains' of spectral players over some baseline dataset $\mathcal{X}$ and decompose the 'gains' among them. In this paper, we use image classifiers to conduct experiments. The decision risk function design for the case when models output embeddings is provided in supplementary material.

There are two candidates for image classifiers: (1) The loss function $\ell : (I, y) \mapsto \mathbb{R}$ can be directly used as the decision risk function $R : (I, y) \mapsto \mathbb{R}$ by simply taking $R(I, y) = \ell(I, y)$, and, (2) the negative prediction probability can also be used as decision risk function by taking $R(I, y) = -\mathcal{P}_{\mathcal{T}}(I, y)$ where $\mathcal{P}_{\mathcal{T}} : (I, y) \mapsto [0, 1]$ measures the probability with respect to label y and $\mathcal{T}$ is a temperature parameter if we choose 'soft' Softmax (Hinton et al., 2015; Goodfellow et al., 2016). In this paper, we choose the latter probability measure with $\mathcal{T} = 1$ as the decision risk function.

[5]We use a *strong efficiency* axiom statement which implies the usual *efficiency* axiom.



(a) googlenet (b) resnet18 (c) efficientnet_v2

Figure 6: This is an application showcase (A3) in research demonstrating how the robustness of feature representations of models evolves with respect to training epochs. We train 'googlenet', 'resnet18' and 'efficientnet_v2' with 120 epochs, SGD optimizer and learning rate 0.0025. The training dataset is *Caltech*101. The spectral importance distributions are measured for every 5 epochs using **I-ASIDE**. The result suggests that models have two learning stages: (1) Learning the feature representations dominated by low-frequency components at an earlier stage and (2) refining by learning the feature representations associated with high-frequency components at later stage. This result from **I-ASIDE** also echos with the major claims from prior works: (1) The results from F-Principle (Xu et al., 2019a; Luo et al., 2019) and (2) the standard accuracy of visual models is at odds with robustness (Tsipras et al., 2018; Ilyas et al., 2019). The full results are provided in supplementary material.

3.4 Spectral coalition filtering

Spectral coalition vector: A spectral coalition vector $C(\widetilde{\mathcal{I}}) = (C_i(\widetilde{\mathcal{I}}))_{i=0}^{M-1}$ is an indicator function for some spectral coalition $\widetilde{\mathcal{I}}$ with dimension M such that:

$$C_i(\widetilde{\mathcal{I}}) := \begin{cases} 1, & \text{if } \mathcal{I}_i \in \widetilde{\mathcal{I}}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Coalition filtering: Let $\mathscr{F}$ be Discrete Fourier transform (DFT) operator and $\mathscr{F}^{-1}$ be inverse DFT (IDFT) operator[6] (Tan & Jiang, 2018). We take some spectral coalition $\widetilde{\mathcal{I}}$ and crop the frequency components not present in $\widetilde{\mathcal{I}}$ by using a channel-wise plane comb filter on frequency-domain. Let $H_{\widetilde{\mathcal{I}}} = (H_{\widetilde{\mathcal{I}}}(m,n))_{(m,n)\in[0,N-1]\times[0,N-1]}$ be the filter transfer function for some spectral coalition $\widetilde{\mathcal{I}}$ such that:

$$H_{\widetilde{\mathcal{I}}}(m,n) = \begin{cases} 1, & \text{if the frequency point (m,n) in } \widetilde{\mathcal{I}}, \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $N \times N$ is the dimension of image. We define an operator '$\star$' to represent the filtering by:

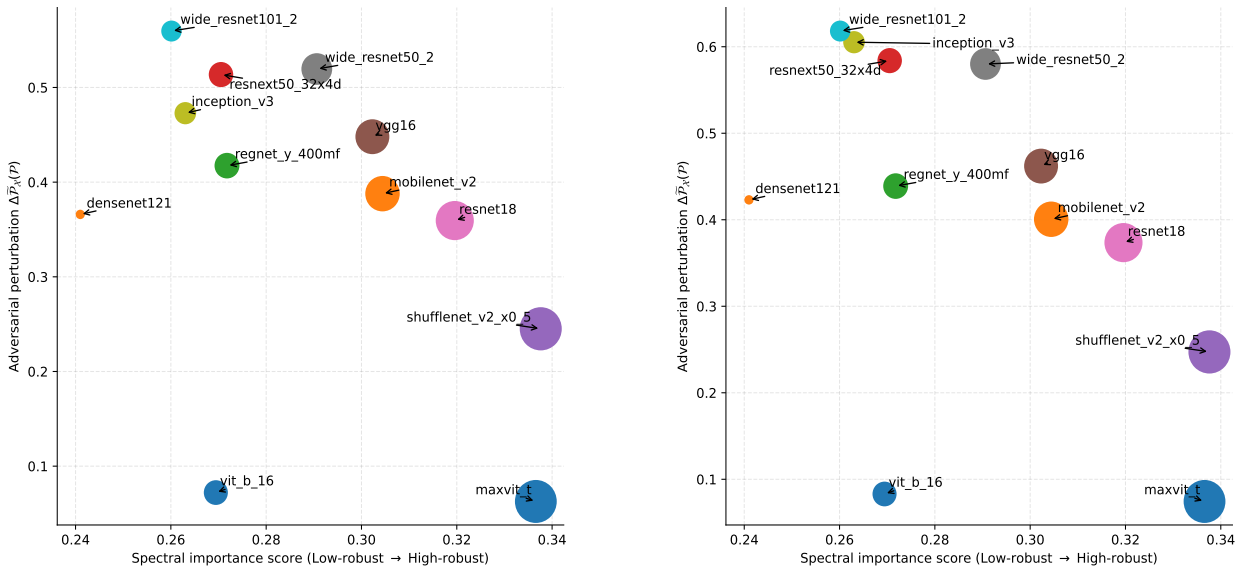
$$\widetilde{I}_{\widetilde{\mathcal{I}}} = I \star \widetilde{\mathcal{I}} = (\mathscr{F}^{-1}(\mathscr{F}(I_c) \odot H_{\widetilde{\mathcal{I}}}))_{c=0}^{2} \tag{3}$$

where '$\odot$' denotes Hadamard product (Kim et al., 2016).

3.5 Decision measure on spectral coalition

Let $R : (I, y) \mapsto \mathbb{R}$ be the decision risk function of some neural network where I is input and y is label. Let $\hat{R}_{\mathcal{X},\widetilde{\mathcal{I}}}(R) : R \mapsto \mathbb{R}$ be the statistical expectation of the decision risk function R on spectral coalition $\widetilde{\mathcal{I}}$

[6]The formal definitions of DFT and IDFT are provided in supplementary material.



(a) FGSM (b) PGD

Figure 7: This is an application showcase (A4) in research demonstrating that there is a correlation between adversarial perturbations and the spectral importance scores using **I-ASIDE**. The adversarial perturbations are measured by the statistical expectations of prediction probability variations (See Figure 4.4). We choose multiple pretrained models (on *ImageNet*) and perform un-targeted FGSM/PGD attacking with $eps = 0.2$. The adversarial samples carry a high proportion of non-robust features (Ilyas et al., 2019). For most models, adversarial perturbations are negatively proportional to spectral importance scores. The circle sizes are proportional to spectral importance scores. It is also notable there are some outliers, and this implies the perturbation-based robustness research can merely capture an aspect of the holistic robustness of models. The full results are provided in supplementary material.

over dataset $\mathcal{X}$. In this paper, the statistical expectation of the decision risks over dataset $\mathcal{X}$ with spectral coalition $\widetilde{\mathcal{I}}$ is defined by:

$$\widehat{R}_{\mathcal{X}}(R, \widetilde{\mathcal{I}}) := \mathop{\mathbb{E}}_{I,y\sim\mathcal{X}} \Big(R(I \star \widetilde{\mathcal{I}}, y) - R(I \star \varnothing, y) \Big). \tag{4}$$

3.6 Decomposition equation

Let $V_{\mathcal{X}}(R) = (V_{\mathcal{X},\mathcal{I}_i}(R))_{i=0}^{M-1}$ be the decomposition of the statistical expectation of some negative decision risk function $-\widehat{R}_{\mathcal{X}}(R, \widetilde{\mathcal{I}})$ and $V_{\mathcal{X},i}(R)$ be the importance of the i-th spectral player $\mathcal{I}_i$ for some neural network $R : (I, y) \mapsto \mathbb{R}$ over some dataset $\mathcal{X}$. The $V_{\mathcal{X}}(R)$ must satisfy the *strong efficiency* axiom by:

$$C(\widetilde{\mathcal{I}})V_{\mathcal{X}}(R) = -\widehat{R}_{\mathcal{X}}(R, \widetilde{\mathcal{I}}) \tag{5}$$

for $\forall\widetilde{\mathcal{I}} \subseteq \mathcal{I}$ and $\widetilde{\mathcal{I}} \neq \varnothing$ where $C(\widetilde{\mathcal{I}})$ is the spectral coalition vector. It is not difficult to show that the equation (5) satisfies: (1) '*symmetry* axiom', (2) '*linearity* axiom' and (3) '*null player* axiom'[7].

Let $C_{(2^M-1)\times M} := (C(\widetilde{\mathcal{I}}))_{\widetilde{\mathcal{I}}\subseteq\mathcal{I}\wedge\widetilde{\mathcal{I}}\neq\varnothing}$ be a collection of every coalition vector $C(\widetilde{\mathcal{I}})$ $(\widetilde{\mathcal{I}} \neq \varnothing)$. We refer C as the '*spectral coalition matrix*'[8]. Let $\widehat{R}_{\mathcal{X}}(R) := (\widehat{R}_{\mathcal{X}}(R, \widetilde{\mathcal{I}}))_{\widetilde{\mathcal{I}}\subseteq\mathcal{I}\wedge\widetilde{\mathcal{I}}\neq\varnothing}$ be a collection of the statistical expectation of every decision risk with respect to spectral coalitions. The equation (5) can be rewritten into the form of matrix:

$$CV_{\mathcal{X}}(R) = -\widehat{R}_{\mathcal{X}}(R). \tag{6}$$

[7]The proof that the decomposition equation satisfies the distribution axioms is provided in supplementary material.
[8]An example of a spectral coalition matrix with $M = 3$ is provided in supplementary material.

Hence the spectral importance decomposition can be represented as:

$$V_{\mathcal{X}}(R) = -(C^TC)^{-1}C^T\widehat{R}_{\mathcal{X}}(R) = -(C^TC)^{-1}C^T\left(\widehat{R}_{\mathcal{X}}(R,\widetilde{\mathcal{I}})\right)_{\widetilde{\mathcal{I}}\subseteq\mathcal{I}\wedge\widetilde{\mathcal{I}}\neq\varnothing} \tag{7}$$

where C^TC is an invertible symmetric real matrix[9].

3.7 Spectral importance score

We summarize the derived spectral importance distributions into scalars by comparing against a feature representation robustness prior: The learned feature representations dominated by low-frequency components are more robust than those feature representations dominated by high-frequency components. We thus devise a hand-crafted weighting vector $\boldsymbol{\beta} := (\beta^k)_{k=0}^{M-1}$ as a 'prior' emphasizing low-frequency components to summarize the robustness by summing with weighting. We also scale scores to $[0, 1]$ against random decisions. The empirical spectral importance score $S_{\mathcal{X}}(R)$ of some model R over some baseline dataset $\mathcal{X}$ is given by[10]:

$$S_{\mathcal{X}}(R) := \left|\frac{\boldsymbol{\beta}^{*T}V^*_{\mathcal{X}}(R) - \eta}{1-\eta}\right| \tag{8}$$

where $\beta \in (0, 1)$, $\boldsymbol{\beta}^* = \frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2}$, $\eta = \frac{1}{M}\frac{||\boldsymbol{\beta}||_1}{||\boldsymbol{\beta}||_2}$ and $V^*_{\mathcal{X}}(R) = \frac{V_{\mathcal{X}}(R)-\min V_{\mathcal{X}}(R)}{||V_{\mathcal{X}}(R)-\min V_{\mathcal{X}}(R)||_1}$ is a normalized probability distribution. We choose $\beta = 0.75$ in this work.

3.8 Estimation of error upper bound

Suffering from the limit of computational resources, the statistical expectation is evaluated by using Monte Carlo sampling. We analyze the error bound by taking K samples. Let $V^*_{\mathcal{X}}(R)$ and $\widehat{R}^*_{\mathcal{X}}(R)$ be the estimations of the spectral importance and decision risks using Monte Carlo. The error bound with L_1 norm is given by:

$$\epsilon_{\mathcal{X}}(V;R) = \mathop{\mathbb{E}}_{\mathcal{X}}||V^*_{\mathcal{X}}(R) - V_{\mathcal{X}}(R)||_1 \leqslant ||(CC^T)^{-1}C||_1 \cdot M \cdot \left\{\frac{Var(\widehat{R}_{\mathcal{X}}(R))}{K}\right\}^{\frac{1}{2}}$$

where K gives the number of samples, $Var(\widehat{R}_{\mathcal{X}}(R))$ gives the variance of the decision risk functional $\widehat{R}_{\mathcal{X}}(R)$ of model R and M is the number of spectral regions. The full derivation is provided in supplementary material. In our experiments, we empirically choose $K = 200$ due to this error upper bound. For example, evaluating resnet18 on *ImageNet* with $K = 200$, the L_1 norm of the distribution difference is less than 0.007.

4 Applications

We showcase multiple applications to demonstrate the potential uses of **I-ASIDE**[11]: (A1) Quantifying model global robustness by examining spectral importance distributions, (A2) understanding the learning behaviours of image models on the datasets with supervision noise, (A3) investigating the learning dynamics of models from the perspective of the evolution of feature representation robustness in optimization and (A4) understanding the adversarial vulnerability of models by examining spectral importance distributions.

4.1 Showcase A1: Quantifying model global robustness

We measure the robustness by examining the spectral importance distributions or the spectral importance scores.

Spectral importance distribution: The experiments in Figure 3 showcase the measured spectral distributions of multiple models with multiple datasets. The results show that the spectral importance of the learned

[9]The proof of the invertibility of the C^TC matrix is demonstrated in supplementary material.
[10]The deduction of this formula is provided in supplementary material.
[11]The core code implementation is provided in supplementary material.

feature representations of trained models is anisotropic. Furthermore, models trained on larger training datasets exhibit higher robustness.

Numerical comparison: The experiment (a) in Figure 4 showcases the application of numerically comparing model robustness. The results correlate with the adversarial perturbation experiments in Figure 7 in which we measure the prediction probability variations between clean samples and adversarial samples for given labels with un-targeted FGSM/PGD attacking (Szegedy et al., 2013; Moosavi-Dezfooli et al., 2016; Goodfellow et al., 2014; Madry et al., 2017). The result implies that the numbers of trainable parameters seem not to play a crucial aspect on model robustness.

Visualizing by projection: The experiment (b) in Figure 4 showcases the projection of spectral importance distributions. The projection is performed by using t-SNE (Hinton & Roweis, 2002).

Interestingly, the very light model *shufflenet_v2_x0_5* (1.4M weights) (Zhang et al., 2018; Ma et al., 2018; Shao et al., 2021) has a high robustness and Tang et al. also report similar observations. Our later adversarial perturbation experiments in Figure 7 can further affirm this. The speculation behind is that the channel shuffle operations in *shufflenet* impose the 'architectural bias' to encourage the channel-wise feature interactions, and can implicitly guide neural networks to ignore the features sensitive to such operation-wise 'perturbations'.

4.2 Showcase A2: Understanding how models respond to supervision noise

The experiments in Figure 5 showcase the investigation of how models respond to various label noise levels in the training process. We conduct the experiment using the *Caltech101* dataset. We create noisy-label datasets from the clean *Caltech101* by randomly re-assigning a proportion (referred to as'label noise level' or 'supervision noise level') of clean labels. We vary the label noise level from 0 to 1 with stride 0.1. The results show that the models trained with higher label noise levels tend to use a higher proportion of non-robust features.

4.3 Showcase A3: Investigating how feature representations are learned in training

The experiments in Figure 6 showcase how the feature representations learned by models evolve with respect to epochs in training process. The learning dynamics of models exhibit two stages. In the first stage, models readily learn more robust feature representations. In the second stage, models learn more non-robust feature representations to achieve higher accuracy. This result gives an insight into designing better optimization goals by introducing robustness penalty towards more robust models.

4.4 Showcase A4: Understanding adversarial vulnerability

Adversarial samples carry a high proportion of non-robust features (Su et al., 2018; Ilyas et al., 2019; Bai et al., 2021; Tsipras et al., 2018). Intuitively, the models with high spectral importance scores are less likely susceptible to adversarial perturbations. In Figure 7, we show how FGSM/PGD (Yuan et al., 2019; Madry et al., 2017; Akhtar & Mian, 2018; Chakraborty et al., 2018) adversarial perturbations correlate with the spectral importance scores by **I-ASIDE**. The full results are provided in supplementary material. We measure the prediction probability variations between the prediction probabilities of clean samples and adversarial samples. We use the 'smooth' version of Softmax function with temperature $\mathcal{T}$ to convert logits into probabilities (Hinton et al., 2015; Goodfellow et al., 2016; Jang et al., 2016).

Experimental method: Let I be some clean image and I^* be the corresponding adversarial sample. Let $\mathcal{P}_\mathcal{T} : (I, y) \mapsto [0, 1]$ be some neural network which outputs the probability with respect to some category y by using Softmax with the temperature parameter $\mathcal{T}$. We define the adversarial perturbation $\Delta\widetilde{\mathcal{P}}_\mathcal{X}(\mathcal{P})$ as $\mathop{\mathbb{E}}_{I,y\sim\mathcal{X}} |\mathcal{P}_\mathcal{T}(I, y) - \mathcal{P}_\mathcal{T}(I^*, y)|$ over some dataset $\mathcal{X}$ where (I, y) denotes samples. We set $\mathcal{T} = 2$ for both FGSM and PGD attacking. We plot the prediction perturbations against spectral importance scores.

The results in Figure 7 and the full experiment in supplementary material shows that the adversarial perturbations are negatively correlated with spectral importance scores. The outliers remind us: Spectral importance distributions can merely reflect one aspect of the robustness of models.

5 Limitations

I-ASIDE provides a unique insight into the interpretability of image models from the feature robustness on spectral perspective. Yet it remains noteworthy that spectral perspective can merely reflect one aspect of the holistic view of model robustness. For example, carefully crafted malicious adversarial perturbations on low-frequency components can fool neural networks (Luo et al., 2022; Liu et al., 2023; Maiya et al., 2021). This further implies the complexity of this research topic. Also, as the spectral energy follows power law alike distribution, measuring the expectations of spectral contributions on decisions on high-frequency components poses a challenge. As such, **I-ASIDE** does not use sampling based approach on spectral coalitions to avoid inaccurate results and thus suffers from the computation cost by $\mathcal{O}(2^M)$. Fortunately, we do not need high spectral resolution to analyze model robustness problem.

6 Conclusions

This research is motivated by the link between feature representation robustness and the anisotropy of the spectral structures of natural images, in tandem with the insight that neural networks are parameterized non-linear signal filters. **I-ASIDE** investigates how the 'neural signal filter' respond to signals on spectrum, and quantifies the network transfers by using axiomatic spectral decomposition approach. Our work provides a unique insight into deep learning research and enables a considerable number of applications as we have demonstrated.

References

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Robert J Aumann and Lloyd S Shapley. *Values of non-atomic games*. Princeton University Press, 2015.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.

Paul R Halmos. *Measure theory*, volume 18. Springer, 2013.

Sergiu Hart. Shapley value. In *Game theory*, pp. 210–216. Springer, 1989.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 131–138, 2019.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Jiyuan Liu, Bingyi Lu, Mingkang Xiong, Tao Zhang, and Huilin Xiong. Low frequency sparse adversarial attack. *Computers & Security*, 132:103379, 2023.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15315–15324, 2022.

Tao Luo, Zheng Ma, Zhi-Qin John Xu, and Yaoyu Zhang. Theory of the frequency principle for general deep neural networks. *arXiv preprint arXiv:1906.09235*, 2019.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Shishira R Maiya, Max Ehrlich, Vatsal Agarwal, Ser-Nam Lim, Tom Goldstein, and Abhinav Shrivastava. A frequency perspective of adversarial robustness. *arXiv preprint arXiv:2111.00861*, 2021.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. *Advances in neural information processing systems*, 27, 2014.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems*, 29, 2016a.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*, 2016b.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding neural networks via feature visualization: A survey. *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 55–76, 2019.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *international conference on machine learning*, pp. 2847–2854. PMLR, 2017.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Alvin E Roth. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 631–648, 2018.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Lizhe Tan and Jean Jiang. *Digital signal processing: fundamentals and applications*. Academic Press, 2018.

Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, et al. Robustart: Benchmarking robustness on architecture design and training techniques. *arXiv preprint arXiv:2109.05211*, 2021.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Yusuke Tsuzuku and Issei Sato. On the structural sensitivity of deep convolutional networks to the directions of fourier basis functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 51–60, 2019.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8684–8694, 2020.

Eyal Winter. The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054, 2002.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019a.

Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *International Conference on Neural Information Processing*, pp. 264–274. Springer, 2019b.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.

Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pp. 2528–2535. IEEE, 2010.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, 2018.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.