The Myth of Robust Classes: How Shielding Skews Perceived Stability

Ananyapam De
TU Clausthal
ananyapam.de@tu-clausthal.de

Benjamin Säfken
TU Clausthal
benjamin.saefken@tu-clausthal.de

Abstract

We introduce an information-geometric framework for understanding class robustness of deep vision classifier models, leveraging the manifold structure of a classifier's predictive distribution. Our approach quantifies class robustness using Fisher-Rao (FR) Margins, which measure the distance from an input to it's decision boundaries minimizing a given loss function. Furthermore, this reveals class shielding effects, where unintended intermediate classes impede transitions between source and target classes, offering insights into model vulnerabilities. We hypothesize that classes deemed "robust" often exhibit high shielding frequency, suggesting apparent robustness can stem from large input-volume mapping rather than a faithful understanding of the decision boundary. We perform experiments on CIFAR-10 test images to understand the geometry of the pre-trained classifier, including class transitions, shielding phenomena, and differential class stability.

1. Introduction

Deep neural networks have achieved remarkable performance on a wide range of computer vision tasks, but their black-box nature motivates continuing work on interpretability and trustworthy decision making (XAI) [16]. A large body of work in explainable vision focuses on feature-level attributions—saliency maps [30], Integrated Gradients [32], and class activation map variants such as Grad-CAM [29], Grad-CAM++ [7], Ablation-CAM [11], Score-CAM [36], LayerCAM [15], Eigen-CAM [4], and KPCA-CAM [18]. While these methods illuminate which image regions drive predictions, they often treat the model's output only through local, feature-centric lenses and can overlook shifts in the entire predictive distribution that underlie robustness failures [10, 34].

Adversarial vulnerability has exposed precisely this gap: imperceptible input perturbations can induce large, systematic changes in predicted probabilities and cause misclassification [13, 25, 33]. Both white-box attacks that exploit gradient information (e.g., FGSM, BIM, PGD, Carlini–Wagner) and black-box strategies that rely on transferability or query-based gradient estimation demonstrate that robustness must be considered at the level of the model's predictive distribution and decision boundaries [2, 6, 13, 20, 22, 27]. Attempts to defend by masking gradients or ad-hoc preprocessing have in turn been shown vulnerable to adaptive attacks, motivating principled defenses such as adversarial training and certification via randomized smoothing or convex relaxations [3, 9, 22, 39].

A key limitation of many robustness analyses is their reliance on Euclidean or pixel-space distances (e.g., classical ℓ_2 / ℓ_∞ norms used in DeepFool and other attacks) to measure perturbation size and proximity to decision boundaries [13, 25]. Such distances do not respect the intrinsic, nonlinear geometry induced by a classifier's predictive distribution. Information geometry equips the space of model outputs with the Fisher Information Matrix as a Riemannian metric, one measures meaningful (locally parameter-invariant) distances on the statistical manifold of predictive distributions [1, 26]. Prior work has explored Fisher-based attacks (e.g., OSSA) but has typically considered single-step or localized perturbations rather than complete geometric paths [40].

In this paper we propose a framework that quantifies robustness by measuring Fisher–Rao margins: distances on the statistical manifold induced by the model's predictive distribution that indicate how far an input must move (in information-geometric terms) before the classifier's decision changes toward a target class. The natural gradient is the steepest-descent direction for a loss when distances between parameter values are measured by the local KL (Fisher) metric, i.e. it minimizes the first-order change in the loss subject to a fixed infinitesimal KL constraint (local quadratic approximation). Unlike purely Euclidean measures, FR margins reflect the intrinsic distortion of the predictive distribution and therefore emphasize perturbations that meaningfully alter the model's probabilistic beliefs [14, 38].

Our approach differs from other information-theoretic or distributional XAI techniques (e.g., Information-Bottleneck saliency [41], mutual-information based attributions [12],

and virtual adversarial training [24]) since we explicitly use an information geometric view of predictive distributions. We also position our work with respect to distributional robustness and certification literature [9, 31] and to Shapley-value style explanations adapted for vision [5, 21, 28].

We summarize our contributions below:

- We introduce an information-geometric framework to quantify model robustness for deep vision classifiers via FR margins, computed efficiently by following normalized natural-gradient trajectories on the statistical manifold and identifying boundary points that induce class transitions.
- We empirically demonstrate *class shielding effects*, showing how intermediate classes can block direct paths and thereby influence adversarial transition patterns.
- We validate our framework on CIFAR-10 test images to understand class stability with implications for robust model design.

2. Fisher-Rao Margins

The *Fisher Information Matrix* (FIM) at an input point $\mathbf{x}$ is central to this framework. It precisely quantifies how sensitively the model's output probability distribution $p(y|\mathbf{x};\theta)$ reacts to infinitesimal input perturbations, thereby defining a Riemannian metric on the input space $\mathbb{R}^d$ [1, 26]. Formally, for a model with fixed weights θ, the FIM $\mathbf{G}_\mathbf{x}$ is given by:

$$\mathbf{G}_\mathbf{x} = \mathbb{E}_{y|\mathbf{x};\theta}[(\nabla_\mathbf{x} \log p(y|\mathbf{x};\theta))(\nabla_\mathbf{x} \log p(y|\mathbf{x};\theta))^\top]$$

For classification, where $p_c(\mathbf{x};\theta)$ is the probability of class c and $J(y,\mathbf{x};\theta) = -\log p(y|\mathbf{x};\theta)$, the FIM is explicitly:

$$\mathbf{G}_\mathbf{x} = \sum_{c=1}^{C} p_c(\mathbf{x};\theta)[\nabla_\mathbf{x} J(c,\mathbf{x};\theta)][\nabla_\mathbf{x} J(c,\mathbf{x};\theta)]^\top$$

This FIM is distinct from parameter-focused formulations in [24], and acts as a local metric tensor on the statistical manifold of model outputs. It avoids distortions often seen with the empirical Fisher approximation [19]. Geometrically, if $\mathbf{z} = f(\mathbf{x};\theta)$ is the softmax output, $\mathbf{G}_\mathbf{x}$ is the Riemannian metric induced from the FIM $\mathbf{G}_\mathbf{z}$ in the probability space [26]: $\eta^\top \mathbf{G}_\mathbf{x} \eta = \eta^\top \mathbf{J}_f^\top \mathbf{G}_\mathbf{z} \mathbf{J}_f \eta$, where $\mathbf{J}_f$ is the Jacobian. This implies that the geodesic distance in the lower-dimensional probability space is no larger than in the high-dimensional input space, extending the excessive linearity explanation [13] to networks with smooth activations [8].

Using the FIM, we extend Euclidean robustness metrics like DeepFool [25] to the information-geometric setting. The infinitesimal Fisher-Rao distance ds is defined by $(ds)^2 = d\mathbf{x}^\top \mathbf{G}_\mathbf{x} d\mathbf{x}$. For a path $\gamma(t)$, the total Fisher-Rao distance is:

$$D_{FR}(\mathbf{x}_0, \mathbf{x}_1) = \int_0^1 \sqrt{\dot\gamma(t)^\top \mathbf{G}_{\gamma(t)} \dot\gamma(t)}$$

The Fisher-Rao margin for input $\mathbf{x}$ with true label y can then be approximated as:

$$\text{margin}(\mathbf{x}, y) = \min_{\mathbf{x}'} \{D_{FR}(\mathbf{x}, \mathbf{x}') \mid \arg\max_c p_c(\mathbf{x}';\theta) \neq y\}.$$

This provides a robust, sample-specific measure of robustness that accounts for the model's sensitivity in its output distribution.

As exact margin computation is NP-hard [31] and explicitly requires solving the geodesic boundary-value problem, we employ an iterative procedure based on natural gradient descent [1] to approximate these margins. For a target class k, we seek $\mathbf{x}'$ where $L_k(\mathbf{x}) = \log p_y(\mathbf{x};\theta) - \log p_k(\mathbf{x};\theta)$ becomes zero. Starting from $\mathbf{x}_0$, we compute the natural gradient direction v_k by solving

$$(\mathbf{G}_{\mathbf{x}_0} + \lambda I) v_k = \nabla_\mathbf{x} L_k(\mathbf{x})$$

using CG [23]. The λI term ensures numerical stability for potentially ill-conditioned FIMs [19]. The step magnitude is $\alpha = \sqrt{g^\top v_k}$ and we take a unit Fisher-length step $\delta = v_k/\alpha$. We accumulate α which is interpretable as cumulative linearized logit change. In the continuous, exact-solve limit, this equals the total change in the logit-difference L_k from start to boundary and therefore is a natural and consistent proxy for the required logit change to reach misclassification.

A particularly insightful aspect revealed by this iterative procedure is *class shielding*. As we navigate the manifold from a source class towards a target class k, the path may unexpectedly first encounter the decision boundary of an *intermediate, unintended class* $j \neq k$. This provides us the direct empirical evidence of the complex structure of the model's decision landscape. While the concept of boundaries under the Fisher metric has been explored in single step adversarial attacks like OSSA [40], our method uniquely reveals these intermediate class interactions using iterative procedures.

Direct FIM computation and inversion are prohibitive for high-dimensional inputs ($d \gg 10^4$). We instead compute the FIM-vector product $\mathbf{G}_\mathbf{x}\eta = \mathbb{E}_{y|\mathbf{x};\theta}[(g_y^\top \eta) g_y]$ efficiently via Monte Carlo sampling from $p(y|\mathbf{x};\theta)$ using the alias method [37], typically with $C/5$ samples.

The an approximation of overall Fisher-Rao margin for an individual input $\mathbf{x}$ is the minimum of d_k across all possible target classes $k \neq y_0$. By aggregating these margins across numerous samples within each class, our framework enables the quantitative evaluation of *class robustness*, identifying which classes exhibit larger (more resilient) or smaller (more vulnerable) average Fisher-Rao margins.

This provides a powerful, quantitative metric for assessing differential robustness across the entire spectrum of classes learned by the deep vision model, aligning with recent work on class-level analysis in continual learning contexts [35].

Algorithm 1 Iterative Fisher-Rao Margin Computation

Require: Model $p(y \mid \cdot)$, initial input $\mathbf{x}_0$, current predicted label y_0, target class k
Ensure: Approximate margin d_k, boundary point $\mathbf{x}'_k$ (for target k)

$d_k \leftarrow 0$, $\mathbf{x}_{cur} \leftarrow \mathbf{x}_0$
$L_k(\mathbf{x}) = \log p_{y_0}(\mathbf{x}; \theta) - \log p_k(\mathbf{x}; \theta)$
while $\arg\max_i p_i(\mathbf{x}_{cur}; \theta) = y_0$ **do**
  $g \leftarrow \nabla_{\mathbf{x}} L_k(\mathbf{x})$ at $\mathbf{x}_{cur}$
  $\mathrm{G}(\eta) = \mathbb{E}_{y|\mathbf{x}_{cur};\theta}[(g_y^\top \eta) g_y]$
  Solve $(\mathrm{G} + \lambda I)v = g$ for v via CG
  $\alpha \leftarrow \sqrt{g^\top v}$
  $\delta \leftarrow v/\alpha$
  $\mathbf{x}_{cur} \leftarrow \mathbf{x}_{cur} + \delta$
  $d_k \leftarrow d_k + \alpha$
end while
return d_k, $\mathbf{x}_{cur}$

3. Experiments

We empirically study FR margins and the *shielding* phenomena that arise when following natural gradient trajectories on a pretrained image classifier. For each input image $\mathbf{x}_0$ whose prediction is y_0, we run Algorithm 1 toward every target class k. For every attempted path (a single $(\mathbf{x}_0, y, k)$ run) we store the cumulative linearized change in the logit difference at the endpoint and the terminal predicted class. From the full corpus we extracted and analysed the subset of 800 image runs in which the model's prediction changed, yielding $N = 3{,}659$ changed-path records spanning the ten classes.

We first study the path matrix Figure 1 which gives us an overview of individual class robustness. This gives us an insight that suggests classes containing animals and birds are the most vulnerable. A possible reason is their relatively smaller size, which makes the model's predictions reliable only in close-up images. In contrast, larger objects such as trucks, airplanes, automobiles, ships, and horses can still be identified accurately even from distant images. Notable cross-class transitions include dog→cat, and cat→frog and airplane→bird and bird→airplane, revealing specific adversarial vulnerabilities, hence we study them differently.

We study the FR distances of the paths which change prediction in Figure 2 which reveal that vulnerable classes require smaller fisher distances for successful attacks. This observation is in tandem with [25, 40] who show that such classes require smaller Euclidean perturbations.

We now study the shielding effects for both the cases,

Figure 1. Path matrix showing the percentage of FR paths from each source class (rows) to final predicted class (columns). Values represent the proportion of paths that end up predicting each class when starting from a given source. Diagonal elements indicate self-prediction robustness. Most robust classes (highest self-prediction): ship, truck, airplane, automobile, and horse. Most vulnerable classes (lowest self-prediction): frog, dog, cat, and bird, which are more likely to transition to other classes.

when we try to find a path from the low vulnerability class to a high vulnerability class and vice versa. Figure 3 shows the boxplots of the FR margins of the top 3 frequent shielding classes the for every source-target pair. The shielding analysis reveals that specific classes dominate as barriers. The class truck emerges as the most effective shielding class, demonstrating a 92% shielding frequency, followed closely by ship with 88% frequency and airplane with 72% frequency. The analysis shows that low vulnerability classes (truck, ship, airplane) serve as the most reliable shields, with several source-target combinations achieving 100% shielding ratios - including airplane→deer, airplane→dog, ship→dog, and truck→dog. This is an interesting as well as crucial finding, implying that a large volume of the input region of the model predicts into one of the low vulnerability classes.

These measurements support two structural hypothesis about the classifier's decision boundaries. First, attempts to follow natural-gradient trajectories toward a particular target often terminate in an shielded class and the classes which appear as shields are often the low vulnerability classes. Hence, apparent robustness at the class level can arise for different reasons. A class that exhibits low measured vulnerability may do so because the model accurately

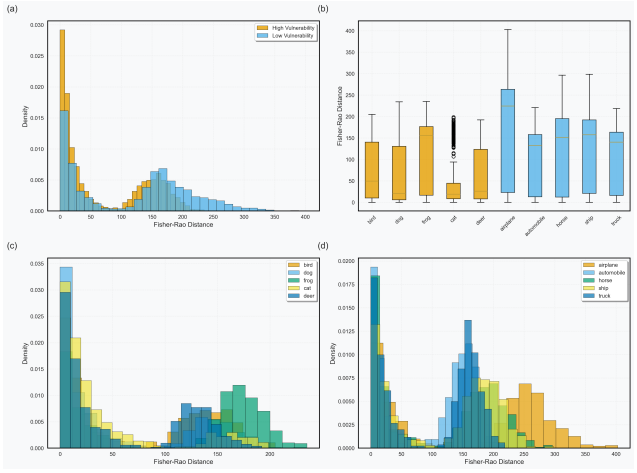

Figure 2. Analysis of Fisher-Rao distances for successful adversarial attacks (changed predictions only) across CIFAR-10 classes. (a) Distribution comparison showing that high vulnerability classes (yellow) require significantly smaller Fisher-Rao distances than low vulnerability classes (blue) for successful class transitions. (b) Box plots by individual class, revealing that high vulnerability classes (bird, dog, frog, cat, deer) have significantly lower median distances compared to low vulnerability classes (airplane, automobile, horse, ship, truck). (c) Individual distance distributions for high vulnerability classes, with cat showing the lowest distances (mean: 41.76) and frog the highest (mean: 111.67). (d) Individual distance distributions for low vulnerability classes, with airplane requiring the largest distances (mean: 162.68) and automobile the smallest (mean: 98.34).

captures its boundary, or because a large volume of input space is mapped to that class by the model. Distinguishing these two mechanisms is crucial for understanding and improving robustness metrics. We hypothesize a "robust" class, automatically makes it into a class demonstrating high shielding frequency. This forces us to revisit the ideas of class robustness, that if certain classes are robust because the model correctly captures their decision boundaries or if the model simply maps a large volume of the input region to these classes making them "seem" robust. Second, the shielding phenomenon is not purely geometric: many ordered pairs are effectively blocked not because the target is extremely far in Fisher length but because other class regions lie between the source and the intended target in geometry. This observation has quite strong implications for robustness measurements and suggests we take into account the topology of the classifier prediction regions as well.

We do note several limitations of this yet ongoing work. Firstly, the experiments till now only use a single pretrained classifier; the accuracy of our claims would require testing this for a larger number of models, so generalisation across models remains an important direction for future work. Recent studies on efficient Fisher estimation and on the shortcomings of empirical Fisher approximations have highlighted the importance of using the true FIM and recognizing the pitfalls of empirical substitutes [17, 19, 35]. Instead we use the approximate FIM to make the computational cost managable for the CIFAR-10 test dataset. Moreover, the efficiency gains in computing the FIM-vector product rely on Monte Carlo sampling with a limited number of samples $(C/5)$, introducing an element of stochasticity and potential inaccuracy into each step of the natural gradient descent. The iterative solver uses a Tikhonov regularization term, and a CG stopping tolerance, and while we verified that the principal qualitative phenomena are robust to modest changes in damping, a comprehensive sensitivity analysis of solver hyperparameters is imperative to reach a stronger conclusion.

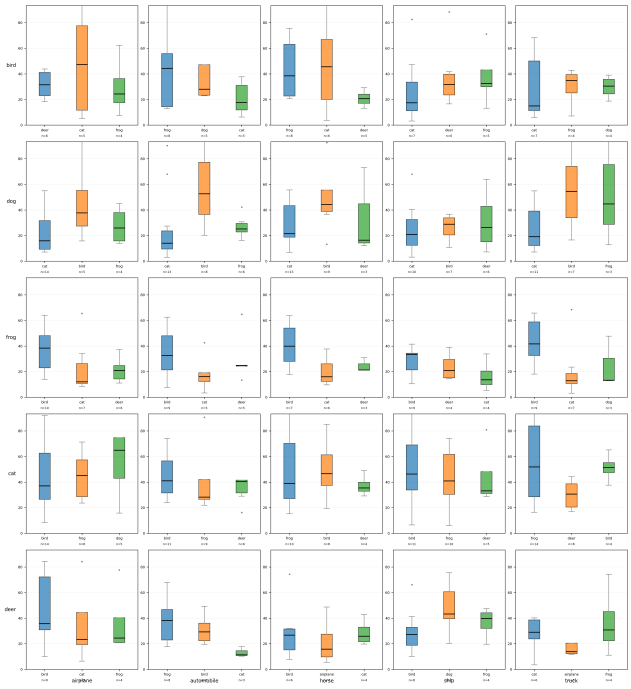

Figure 3. Each cell displays box plots of FR distances for the top 3 most frequent shielding classes that intercept adversarial paths from high-vulnerability source classes (airplane, automobile, horse, ship, truck) to low-vulnerability target classes (bird, dog, frog, cat, deer). The box plots reveal the distribution of decision boundary distances when each shielding class successfully prevents direct adversarial transitions. Higher FR distances indicate stronger shielding effectiveness, as the model requires larger perturbations to bypass these protective classes. Sample sizes are shown below each shielding class name, representing the frequency of each shielding event across multiple adversarial paths.

References

[1] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. 1, 2

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search, 2020. 1

[3] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, 2018. 1

[4] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-CAM: Class activation map using principal components. *arXiv preprint arXiv:2008.00299*, 2020. 1

[5] Huaiguang Cai et al. Cams as shapley value-based explainers. *arXiv preprint arXiv:2501.06261*, 2025. 2

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017. 1

[7] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: Improved visual explanations for deep convolutional networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 1

[8] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus), 2016. 2

[9] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019. 1, 2

[10] Ananyapam De, Anton Frederik Thielmann, and Benjamin Säfken. NODE-GAMLSS: Interpretable uncertainty modelling via deep distributional regression. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*, 2024. 1

[11] Saurabh Desai and Harish G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 972–980, 2020. 1

[12] Itai Gat, Nitay Calderon, Roi Reichart, and Tamir Hazan. A functional information perspective on model interpretation, 2022. 1

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. 1, 2

[14] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features, 2019. 1

[15] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 2021. 1

[16] Vidhya Kamakshi and Narayanan C. Krishnan. Explainable image classification: The journey so far and the road ahead. *AI*, 4(3):620–651, 2023. 1

[17] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 1032–1041. PMLR, 2019. 4

[18] S. Karmani, T. Sivakaran, and L. Prasad. Kpca-cam: Visual explainability of deep computer vision models using kernel pca. *arXiv preprint arXiv:2410.00267*, 2024. 1

[19] Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation for natural gradient descent, 2020. 2, 4

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, 2017. 1

[21] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. 2

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. 1

[23] James Martens. Deep learning via hessian-free optimization. pages 735–742, 2010. 2

[24] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. *arXiv*, 2015. 2

[25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks, 2016. 1, 2, 3

[26] Frank Nielsen. *Geometric Theory of Information*. Springer, 2014. 1, 2

[27] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning, 2017. 1

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. 2

[29] 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 (ICCV)*, pages 618–626. IEEE Computer Society, 2017. 1

[30] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. 1

[31] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training, 2020. 2

[32] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. 1

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. 1

[34] Anton Thielmann, René-Marcel Kruse, Thomas Kneib, and Benjamin Säfken. Neural additive models for location scale and shape: A framework for interpretable neural regression beyond the mean, 2024. 1

[35] Gido M. van de Ven. On the computation of the fisher information in continual learning, 2025. 3, 4

[36] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 24–25, 2020. 1

[37] Jingbo Wang, Wai Tsang, and George Marsaglia. Fast generation of discrete random variables. *Journal of Statistical Software*, 11, 2004. 2

[38] Han Xu, Xiaorui Liu, Yaxin Li, Anil K. Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training, 2021. 1

[39] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy, 2019. 1

[40] Chenxiao Zhao, P. Thomas Fletcher, Mixue Yu, Yaxin Peng, Guixu Zhang, and Chaomin Shen. The adversarial attack and detection under the fisher information metric. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):5869–5876, 2019. 1, 2, 3

[41] Hao Zhou, Keyang Cheng, Yu Si, and Liuyang Yan. Improving interpretability by information bottleneck saliency guided localization. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 1