Adversarial training and attribution methods enable evaluation of robustness and interpretability of deep learning models for image classification

Flávio A. O. Santos,[1,*] Cleber Zanchettin,[1,2,*] Weihua Lei,[3] and Luís A. Nunes Amaral[2,3,4,5,†]

[1]*Centro de Informática, Universidade Federal de Pernambuco, Recife, Pernambuco, 52061080, Brazil*
[2]*Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA*
[3]*Department of Physics and Astronomy, Northwestern University, Evanston, Illinois 60208, USA*
[4]*Northwestern Institute on Complex Systems, Northwestern University, Evanston, Illinois 60208, USA*
[5]*NSF-Simons National Institute for Theory and Mathematics in Biology, Northwestern University, Chicago, Illinois 60611, USA*

(Received 19 September 2023; accepted 19 September 2024; published 21 November 2024)

Deep learning models have achieved high performance in a wide range of applications. Recently, however, there have been increasing concerns about the fragility of many of those models to adversarial approaches and out-of-distribution inputs. A way to investigate and potentially address model fragility is to develop the ability to provide interpretability to model predictions. To this end, input attribution approaches such as Grad-CAM and integrated gradients have been introduced to address model interpretability. Here, we combine adversarial and input attribution approaches in order to achieve two goals. The first is to investigate the impact of adversarial approaches on input attribution. The second is to benchmark competing input attribution approaches. In the context of the image classification task, we find that models trained with adversarial approaches yield dramatically different input attribution matrices from those obtained using standard techniques for all considered input attribution approaches. Additionally, by evaluating the signal-(typical input attribution of the foreground)-to-noise (typical input attribution of the background) ratio and correlating it to model confidence, we are able to identify the most reliable input attribution approaches and demonstrate that adversarial training does increase prediction robustness. Our approach can be easily extended to contexts other than the image classification task and enables users to increase their confidence in the reliability of deep learning models.

DOI: 10.1103/PhysRevE.110.054310

I. INTRODUCTION

Deep learning (DL) approaches are now used regularly in a variety of domains, including drug discovery [1], speech recognition [2], object recognition [3], question and answer [4], machine translation [5], and image description [6]. Strikingly, some published studies report superhuman performance, that is, a performance level exceeding that of human experts [7]. Such claims have created a self-reinforcing cycle of increased popularity, leading to the adoption of deep learning models in ever more areas of our research and development [8–10]. However, as applications move from the mundane—recognizing your friends on social media—to the high stake—self-driving cars [11] or diagnosing SARS-CoV-19 infections from chest x rays [12]—the need to address model fragility [13] and model interpretability becomes increasingly critical.

Pioneering work by Goodfellow and colleagues [14,15] has demonstrated that minute input changes can yield significant changes in the inference process of DL models, an effect they denoted adversarial attacks. Their findings have led to the development of a new approach to model training—adversarial training [15,16].

Despite the importance of this work, adversarial attacks are not the only reason for model failure [17–20]. Even counterintuitive out-of-distribution data can lead to catastrophic model inference failures. For this reason, many researchers argue that in order to address model fragility, we must have model interpretability. Robust and sensible explanations allow users to verify the soundness of a model's performance, used features, and inference assurance. To accomplish the goal of DL model interpretability, researchers have investigated several attribution approaches that aim to estimate the importance of individual features to model inference [21–23].

Attempting to connect these two strands of research, Zhang *et al.* [24] and Tsipras *et al.* [25] have reported that explanations produced by adversarial robust models are more interpretable than explanations generated by standard models. Etmann *et al.* [26] and Ignatiev *et al.* [27] have investigated the connection between adversarial robustness and interpretability from a theoretical perspective. Others have attempted to enhance model generalization by training models with additional objectives that constrain the interpretability or adversarial mask of a given input image [28–32]. Despite these efforts, most of the DL research community has been split between those who focus on increasing model robustness through adversarial training and those who focus on model interpretability as the path to model robustness.

However, even these studies tend to focus on input gradients or a single interpretability method, ignoring the variety

[*] These authors contributed equally to this work.
[†] Contact author: amaral@northwestern.edu

of interpretability methods and the interpretability of inner layers. Moreover, these studies rely on definitions of interpretability that focus on the edges of the foreground object, ignoring the possibility that the interior of the foreground object may contain critical information. In this study, we delve deeper into these questions by combining adversarial robustness and model interpretability in a manner that enables us to answer two important questions quantitatively. The first is whether adversarial approaches impact input attribution distributions across the foreground and background information. The second is whether competing input attribution methods can be benchmarked in relation to human vision.

Within the context of the image classification task, our analyses demonstrate that models trained with similar training approaches yield correlated interpretability maps regardless of model architecture. In contrast, models trained with different training approaches yield uncorrelated interpretability maps. Additionally, by evaluating the signal (typical input attribution of the foreground features) to noise (typical input attribution of the background features) ratio and correlating it to model confidence, we are able to identify the most reliable input attribution approaches and demonstrate that adversarial training does increase prediction robustness.

II. METHODS AND DATA

We focus here on convolutional neural networks (CNN) for image classification since this is a paradigmatic task in DL, specifically the ResNet [33], PreActResNet [34], and Google Inception [35] architectures.

ResNet introduces the concept of residual connection, and instead of learning the mapping from input to output, it learns the residual mapping (i.e., the difference between input and output), which allows training deeper models with hundreds of layers without the vanishing gradient problem. The model PreActResNet is an improvement over the ResNet architecture by incorporating the concept of preactivation residual blocks. In these blocks, the batch normalization and activation functions (i.e., ReLU) are applied before the convolutional layer rather than after. Still, the overall architecture structure remains the same as the original ResNet blocks. Following another direction, the Google Inception Network (also known as the Inception Network or GoogLeNet) proposed the inception modules, which consist of multiple parallel convolutional layers with different filter sizes. These parallel modules allow the model to capture features at multiple scales and resolutions within a single layer.

As shown in Fig. 1(a), we consider four CNN training methods: two classic training—Adam [36] and stochastic gradient descent (SGD) [37,38]—and two adversarial training—fast gradient sign method (FGSM) [15] and projected gradient descent (PGD) [16]. For each CNN model, we obtained input attribution using six widely used interpretability methods—guided Grad-CAM [22], guided backpropagation [23], Grad-CAM [22], saliency [21], integrated gradients [39], and input × gradient [40]. Input attribution approaches produce attribution matrices with the same dimensions as the input. Each attribute matrix value estimates how important the corresponding input feature is to the CNN model's decision. We have used the feature attribution maps as the primary visualization. From this visualization, we can compare if two different trained models distribute their feature importance similarly. We used the original images from the test set of every dataset to preserve the data distribution. We performed no data deformation such as stylize, saturate, or patch shuffle [24].

In order to quantify the similarity of the importance of each input across CNN models, we calculated Spearman's rank correlation of the input attribution matrices across all pairs of CNN models.

We replicated all our analyses on three datasets: 5000 train and 8000 test images with a resolution of 96×96 from STL-10 [41] classified into 10 classes (airplane, bird, car, cat, deer, dog, horse, monkey, ship, and truck), 60 000 train and 10 000 test images with a resolution of 32×32 from CIFAR-10 [42] classified into the same 10 classes, and 21 000 train and 5000 test images with a resolution of 224×224 from RIVAL-10 [43], again classified into the same 10 classes.

A. Adversarial Training

Adversarial training boosts the robustness of machine learning models by intentionally exposing them to adversarial examples during training. This process not only teaches the model to identify and correct vulnerabilities but also promotes the learning of more efficient and robust features. As a result, models become less sensitive to minor variations and more capable of generalizing across the training data set. In this work, we evaluated two adversarial training techniques: the FGSM and PGD. Goodfellow *et al.* [15] proposed the FGSM approach to generate adversarial attacks. It involves calculating the gradient of the loss function with respect to the input vector and obtaining the signs (direction) of each dimension of the gradient vector. The authors argue that the gradient direction is more important than the specific point of the gradient because the space of the input vector is not composed of subregions of adversarial attacks. As suggested, an alternative to make the model robust is to add adversarial samples in the training. In Eq. (1), we present the FGSM adversarial training cost function. Given a standard loss function, J, the input vector x, the label y for input x, and the model parameters θ, FGSM updates the loss as

$$J'(\theta, x, y) = \alpha\, J(\theta, x, y) + (1-\alpha)\, J(\theta, x + \epsilon \odot \text{sign}(\nabla_x J(\theta, x, y)))\,. \tag{1}$$

The first term computes the loss of the model when supplying the original input vector x, while the second step computes the loss of the model when supplied an adversarially generated transformation of x, that is, $x + \epsilon \odot \text{sign}(\nabla_x J(\theta, x, y))$. Notice that ϵ is a random vector with the same dimension as x. As FGSM generates adversarial noise based only on gradient directions, Madry *et al.* [16] proposed the PGD method. They approached adversarial attacks from a min-max perspective, specifically aiming to recognize and defend against the class of attacks they face:

$$\underset{\theta}{min}\ p(\theta), \quad \text{where } p(\theta) = \underset{(x,y)\in D}{\mathbb{E}}\,[\underset{\delta\in S}{max}\ J(\theta, x+\delta, y)]\,. \tag{2}$$

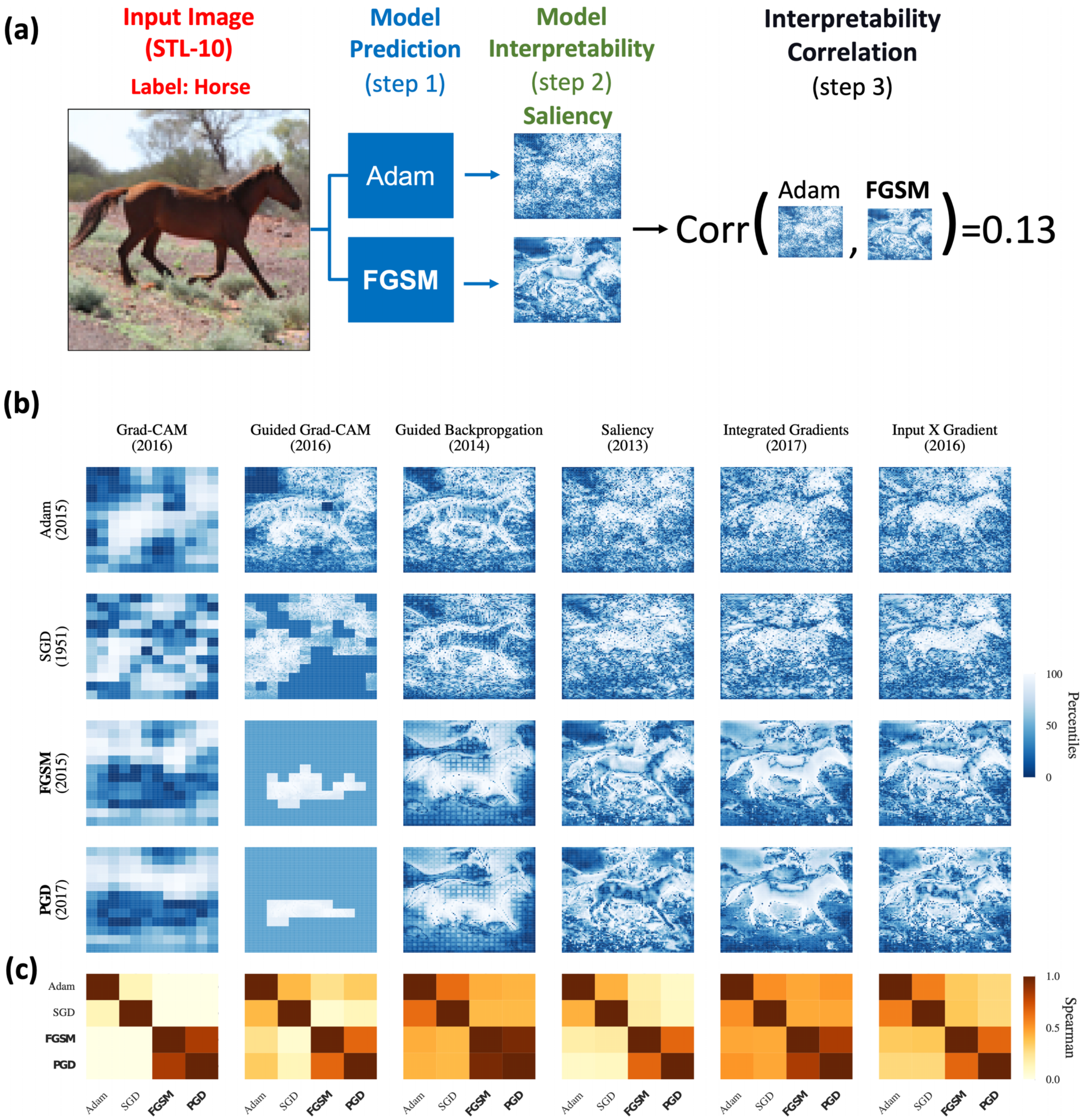

FIG. 1. Do different CNN models focus on different features for determining image classification? (a) Schematic diagram of the experimental pipeline for evaluating the importance of different features on distinct CNN deep learning models. (Step 1) For every selected image in the three datasets considered, we train four learning models that use either classical or adversarial training approaches (Step 1). [Step 2 and (b)] We then use six input attribution approaches to compute the importance of each feature for each of the four learning models. For all attribution approaches, the feature attributions are significantly more similar between adversarially trained models (FGSM and PGD) than between classically trained models. (Adam and SGD). [Step 3 and (c)] Finally, we calculate Spearman's rank correlation of the values in the two input attribution score matrices to determine the similarity of feature attributions obtained from the distinct learning models. The correlation matrices make it visually apparent, for this image, that input attribution scores for every attribution method considered are more similar for models that use the same training approach than for models that use distinct training approaches.

Here, the *max* part pursues adversarial noise that maximizes the loss function J when added to the input vector, while the *min* term seeks to minimize the loss function J, thus making the model robust to max attacks. The authors suggest PGD is a first-order universal attack, as it is the most challenging attack using only first-order information. A limitation of the PGD method is its high computational cost as it requires multiple calculations of the function's gradient to find the max attack.

B. Interpretability methods

Interpretability methods aim to unravel the internal logic of the machine learning models. These methods seek to explain the model's decisions by identifying which input characteristics are most influential. Saliency maps achieve this by calculating the gradients of the model output relative to the input, highlighting the areas that, if changed, would most affect the output. Given an output score i from a model f, it calculates the absolute value of the gradient $\nabla f(x)^i_x$ to obtain the importance of each position in x for the i output for the model f.

Guided backpropagation filters these gradients and focuses only on positive contributions to provide a more unambiguous picture of each aspect of the input that most contributes to the output. Additionally, Grad-CAM can be implemented at the convolutional layers level.

Another gradient-based approach was developed by Sundararajan *et al.* [39]. Their IG method is based on an axiomatic approach. The authors proposed that any adequate interpretability methods must obey two conditions: sensitivity and implementation invariance . Given the input vector of interest x and the baseline vector x', IG estimates the importance I of an input feature by accumulating the gradients of all points on the straight line between the baseline vector and the vector of interest:

$$I_i ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i}\, d\alpha\,. \tag{3}$$

Due to the computational cost of implementing Eq. (3), in practice, one implements an approximate version. In this version, m is the number of points between the baseline vector and the vector of interest, which also represents the number of steps in the Riemann integral approximation

$$I_i^{\text{approx.}} ::= \frac{x_i - x_i'}{m} \times \sum_{k=1}^{m} \frac{\partial F\left(x' + \frac{k}{m} \times (x - x')\right)}{\partial x_i}\,. \tag{4}$$

Although the IG method presents interpretability maps with little noise compared to other purely gradient-based methods, it has two important limitations. First, it requires the (subjective) choice of the baseline [44]. Second, it requires the (subjective) selection of a number m of points on the line that will calculate the gradient to be accumulated.

III. RESULTS

A. Impact of adversarial training on feature importance

In Fig. 1(b), we show the 24 input attribution matrices for the image obtained from each pair of training approach × attribution method. It is visually apparent that input attribution matrices obtained with the same attribution approach and the same type of CNN training method appear more similar. We quantify this hypothesis by calculating Spearman's rank correlation between attribution maps obtained from two different models for the same attribution method [bottom row of Fig. 1(b)]. We use Spearman's rank correlation to compare two attribution maps according to the pixel order of importance, as it can capture any monotonic correlation, not just linear correlations.

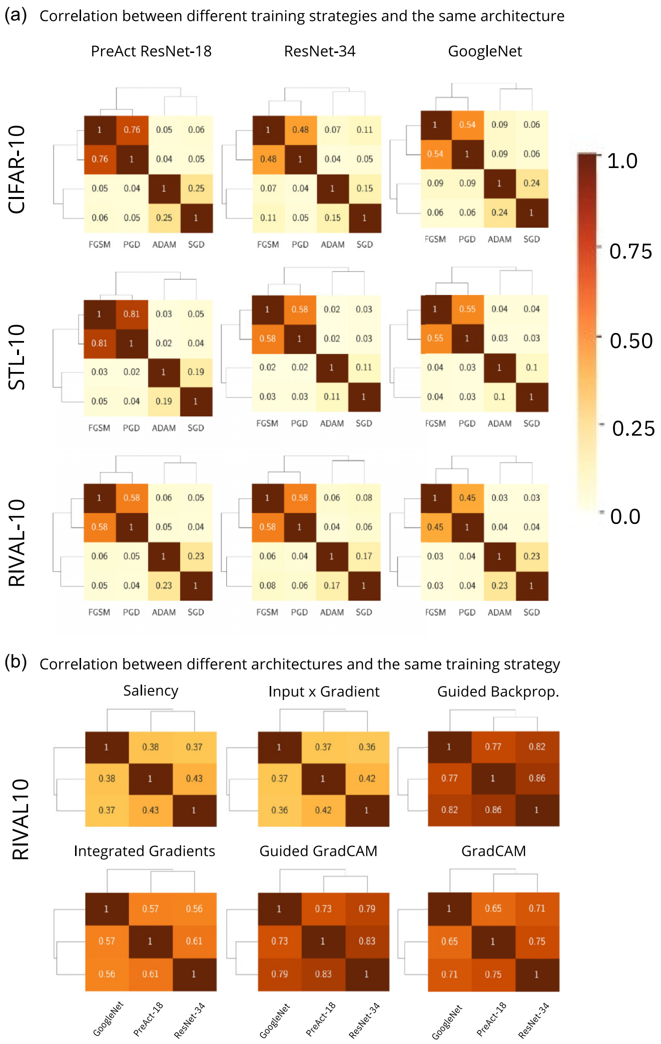

FIG. 2. Input attribution scores for every attribution method considered are more similar for models that are trained using an adversarial approach. (a) We calculate the average (over all images in each of the two datasets considered) of the correlation matrices shown in Fig. 1(c). As suggested by the data in Fig. 1(c), the correlations are stronger between models trained for models that use the same training approach than for models that use distinct training. However, when considering the average across thousands of images, we see that models trained using adversarial approaches yield significantly more correlated input attribution scores than modes using classical training approaches. This finding is consistent with the interpretation that adversarial training decreases the impact of shortcut learning [45]. (b) We compare the correlation matrix between the different PGD models' architecture interpretability. It shows a strong correlation even for different architectures.

Next, we check whether the results shown in Fig. 1 are robust across all images in the three datasets and across the model architecture. In Fig. 2, we show the average of the Spearman correlation across all images from each of the datasets considering six attribution methods. It is visually apparent that across the six attribution methods, (i) adversarial methods produce attribution matrices that are not strongly correlated with those obtained from standard methods, (ii) different adversarial training approaches produce attribution matrices that are more strongly correlated than those obtained from models trained with classic methods, and (iii) different architectures trained using the same strategy produce attribution matrices that are strongly correlated. Thus, we can conclude that adversarial training impacts the importance of features in a significant manner regardless of the attribution method or the model architecture being considered. While prior works have explored the impact of adversarial training in model interpretability, none have addressed this correlation between different adversarial models as our study. This underscores the novelty and significance of our research in pushing the boundaries of understanding in this domain.

B. Relative foreground and background feature importance

Next, we investigate whether adversarial training approaches are able to force the model's training to focus on features of the image that are likely to be important. In order to answer this question, we note that all of the images in these three datasets can be seen as having a foreground (the object being classified) and a background (the rest of the image). This characteristic allows us to pose the question: Do adversarially trained models assign greater importance to foreground pixels than traditionally trained models?

This question is equivalent to asking whether these models are right for the right reasons (RRR) [46–48]. RRR is a necessary property for developing robust machine learning models. RRR models learn which features provide for good performance instead of merely focusing on high performance that could be achieved based on spurious patterns [49]. To investigate this matter, we separated all images in the STL-10 corpus into three groups based on model inference confidence: low ($p < 0.5$), intermediate ($0.5 \leqslant p < 0.9$), and high ($p \geqslant 0.9$). We then selected 40 images from each of the 10 categories for each of the model inference confidence groups. Finally, we manually built foreground masks for those 1200 images (see Fig. 3 for four examples). The masks enable us to separate the foreground from the background when analyzing the attribution matrices obtained for a given set of architecture, training strategy, and attribution method.

Figure 3(d) displays the survival function—defined as one minus the cumulative distribution function—of input attribution values obtained using the IG method, separated by foreground and background, for the PGD (adversarial) and Adam (classical) trained CNN models. It is visually apparent that the distributions decay the fastest (slowest) for the background (foreground) pixels when using the adversarial training method. Comparing adversarial to classical training reveals that the greatest difference is in how adversarial training leads to a reduction of the attribution score associated with background pixels.

To systematically quantify signal and noise, we separately calculate the mean attribution value of input pixels from the foreground and background regions. We then calculate the signal-to-noise ratio for each (image, CNN model, and attribution method) triplet. Figure 4 shows boxplots of signal-to-noise ratios (in decibel units) for the Adam (classical) and PGD (adversarial) trained models according to six attribution methods.

We find that the signal-to-noise ratio is higher for PGD-trained models for all attribution methods except for guided backpropagation. Additionally, for the PGD-trained model, we find that the signal-to-noise ratio increases with inference confidence for all models except, again, for guided backpropagation. Indeed, this increase is statistically significant for the saliency, integrated gradients, and input X gradient attribution methods, suggesting that these three attribution methods may reliably identify RRR models.

Even though the experimental conditions are quite different, our results could potentially be interpreted as contradicting the conclusions of two recent studies. Moayeri *et al.* [43] have reported that adversarially trained models are more sensitive to background information than foreground information when compared to classically trained models. However, they provide no uncertainty estimates for any of the results they display in Fig. 2, raising questions about the statistical significance of the conclusions drawn, especially because they appear not to have tested as many parameter values for adversarially trained ResNets.

Zhang and Zhu [24] used computational experiments to quantify the sensitivity of adversarially trained CNNs to texture and shape distortions. Their study suggests that adversarially trained CNNs are less sensitive to texture distortions and that their robustness appears to be associated with their ability to extract geometric features, such as shape and contours. Our findings (Figs. 3 and 4) align with Zhang and Zhu's conclusions. Our analyses further suggest that adversarially trained CNNs learn to localize the foreground within an image. We speculate that this learning is achieved because of texture differences between the background and foreground, which results in learning the location of the boundary.

C. Revealing the attribution within the hidden layers

The typical DL model comprises several layers [33], but when evaluating adversarial attacks or interpretability maps, researchers focus primarily on the input-to-output or output-to-input path information even though the model's inference relies on the transformations occurring within the intermediate (hidden) layers. To investigate the extent to which attribution scores transform across successive layers, we use the IG attribution method to build input attribution matrices for each intermediate layer. Since models with different architectures but trained using the same approach are already strongly correlated at the input level, we compare here results for a single architecture—PreAct ResNet—but different training approaches [Fig. 5(a)]. For adversarially trained models, the correlations between attribution maps, which start from a high level for the input, continue to increase moderately with successive layers. In contrast, for classically trained models, the correlations between attribution maps increase dramatically with successive layers reaching correlation strengths similar to those observed for the final layers of adversarially trained models [Fig. 5(b)].

Confirming our earlier results that training strategy is more important than model architecture, we find that all three

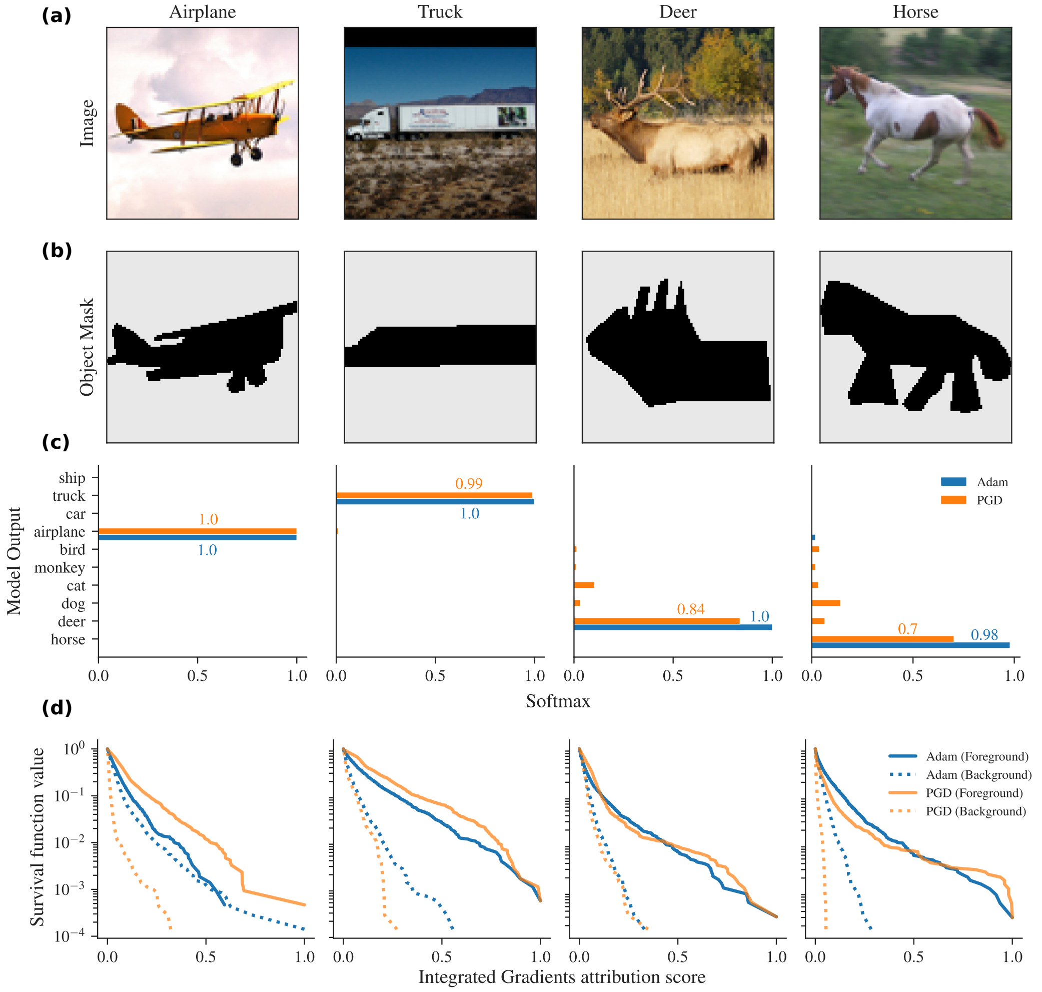

FIG. 3. Deep learning image classification models using adversarial training approaches attribute higher importance to features that fall within an image's foreground. (a) Four selected images from the STL-10 corpus show an airplane, a truck, a deer, and a horse. (b) We manually construct foreground masks for these four images and an additional 1,196 images (not shown). (c) Softmax logits expressing the likelihood of the image belonging to one of the 10 possible classes. (d) Distribution of input attribution scores obtained using the IG method for features within the foreground mask (full lines) and outside the foreground mask (dashed lines) for models trained using the PDG (orange, adversarial) and Adam (blue, classical) training methods. The different decay rates of the survival functions demonstrate that, for these four images, PGD consistently attributes the highest importance to pixels that fall within the foreground mask and the lowest importance to pixels that fall outside the foreground mask.

studied architectures exhibit similar interpretability maps at intermediate and deep layers [Fig. 5(c)]. Under the hypothesis that the interpretability maps obtained with IG capture a model's inference process, our findings suggest that both the intermediate and deep layers of different architectures trained with PGD converge in their focus for inference.

Goodfellow *et al.* [15] suggested that adversarial training is similar to L^2 regularization in a single-layer logistic regression model. Thus, we test the hypothesis that adversarial training produces low weights in the first layers to make the model robust to adversarial changes in the inputs. We apply L^2 regularization to each layer of the model individually and find that penalizing the gradient of the loss with relation to the input vector [50] yields the most robust model against the FGSM attack (Table I). Additionally, we also find that when we penalize only the weights of the first layer, the model achieves adversarial robustness that is at least twice as large as when we regularize the deeper layers, supporting our hypothesis.

TABLE I. Evaluation of the layer-regularization scenario with the CIFAR-10 dataset. In the Layer column, the Conv. L informs that L2 regularization was applied only on layer L. The x grad value suggests that we regularize the input gradient, as proposed in [50].

Model	Layer	Training Accuracy	Testing Accuracy	FGSM
CNN	X grad.	85,91	84,39	51,24
CNN	Conv. 1	88,13	81,51	36,44
CNN	Conv. 2	86,61	82,05	18.18
CNN	Conv. 3	86.86	83.98	15.61
CNN	Conv. 4	87.02	84.77	11.31
CNN	Conv. 5	90.41	86.82	13.54
CNN	Conv. 6	91.91	87.08	15.44

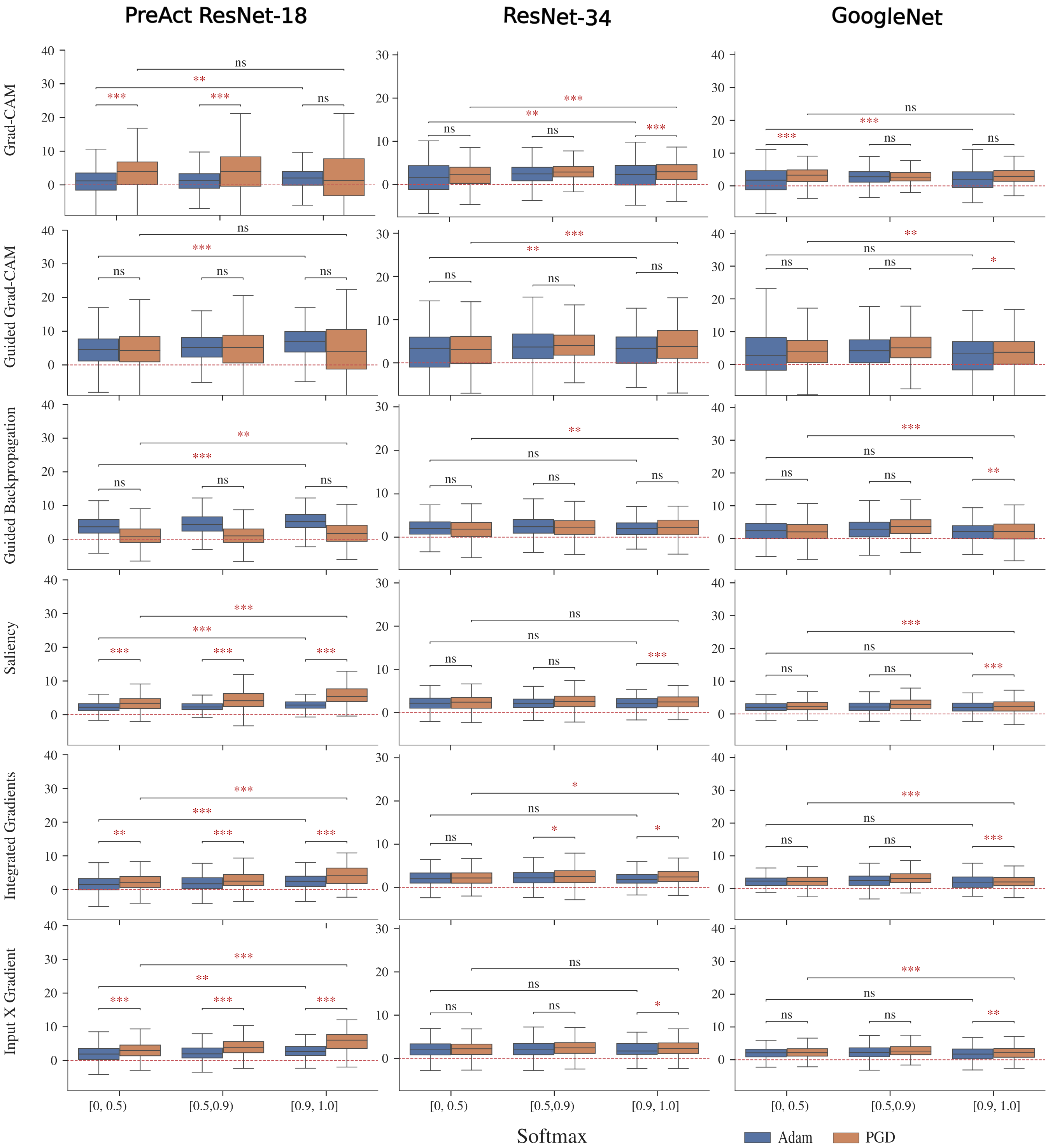

FIG. 4. Deep learning image classification models using adversarial training approaches have greater signal-to-noise ratios for higher confidence inferences. This panel presents the results for three architectures (i.e., PreActResNet-18, ResNet-34, GoogleNet). We define signal as the mean input attribution score of pixels inside the foreground image. We select 1200 images from the STL-10 corpus so that we have the same number of images for each of three inference confidence scores: low confidence (0,0.5), intermediate confidence (0.5, 0.9), and high confidence (0.9, 1). Each of the six panels shows box plots of the signal-to-noise ratios (measured in dB) for the input attribution scores obtained using the labeled input attribution approach and trained either using the Adam (blue, classical training) or PGD (orange, adversarial training) methods. We determine the statistical significance of the differences between cases using a one-tailed student's t-test. We use the convention * for $0.01 < p \leqslant 0.05$, ** for $0.0001 < p \leqslant 0.01$, and *** for $p < 0.0001$). It is visually apparent that the saliency, integrated gradient, and input X gradients input attribution methods display RRR behavior, whereas Grad-CAM, Guided Grad-CAM, and guided backpropagation do not.

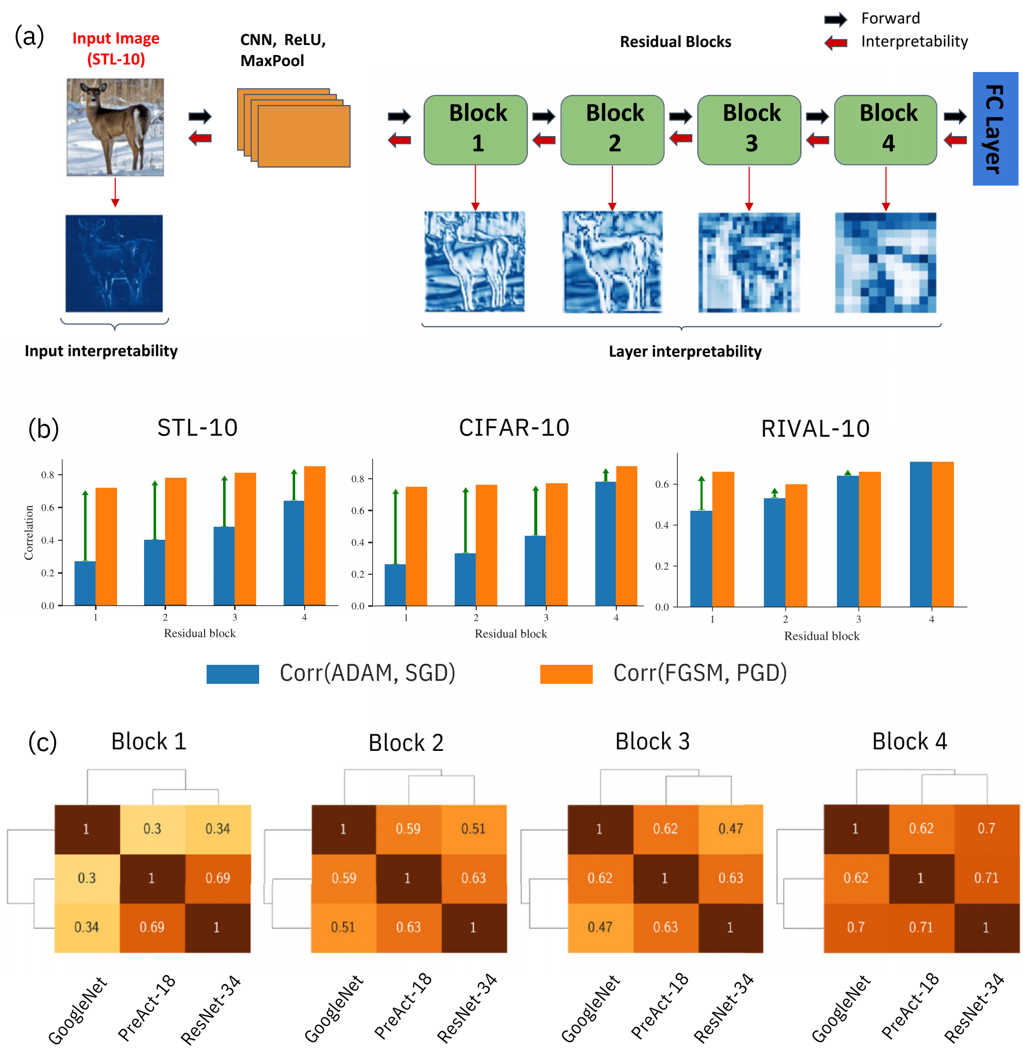

FIG. 5. Adversarial training impacts the most the weights of the first layer. The PreAct ResNet implementation comprises a convolutional layer followed by four residual blocks. Given an input image, the model performs the forward step (black arrows). Next, it back-propagates the interpretations (red arrows) to the first layer. (a) We use the IG method to obtain input attribution score matrices and attribution score matrices for residual blocks 1 to 4. (b) We calculate the average correlation between attribution matrices for models trained using identical approaches (adversarial, orange, or classical, blue). It is visually apparent that adversarially trained models already display high degrees of correlations in their attributions scores for first block values. In contrast, classically trained models only reach large correlations for the final block (the green arrows highlight the decreasing difference in the average correlations with the block index). (c) For each architecture, we compute the PGD model's interpretability for every image of the RIVAL-10 dataset and evaluate the correlation between them. The results show that different architectures trained with PGD on RIVAL-10 produce correlated interpretability maps.

IV. DISCUSSION

Superhuman performance using DL approaches has been reported frequently in various contexts. However, the reasons for such high performance are not well understood. Despite this lack of deep understanding, evidence is building up that some/much of that performance is achieved through shortcut learning—so-called Clever Hans moments. In shortcut learning, the models learn irrelevant (to humans and to the general task) but effective artifacts in the training dataset [12,13,51]. While it has been claimed that this is different from human learning, it may not be so. Indeed, there is wide-ranging literature in behavioral economics about environmentally fit heuristics [52] and what has been denoted as fast thinking [53]. In fact, DL models may be the ultimate builders of dataset-centered successful heuristics. The number of parameters that can be fitted creates the sort of astonishing, but ultimately barren, memorization captured so effectively by Jorge Luis Borges in Funes the Memorious [54]. This brings us to a major concern that potential users in high-stake domains must consider. While heuristic shortcuts can be incredibly effective, they have clear weaknesses and can result in catastrophic failure.

Our study contributes to understanding the conditions under which specific DL models may fail catastrophically. Both adversarial training and input attribution methods have been proposed separately to address the known fragility of DL models. While adversarial training can increase model robustness [15,16], it lacks the transparency to demonstrate whether improvements are not due to other artifacts. By exploiting the strengths of each of these approaches, we learn two very important insights. First, the saliency, IG, and Input X Gradient methods do a good job of capturing the features found to be important by a DL model, but other attribution methods do not. Second, adversarially trained models assign more importance to image foreground features (signal) than to image background features (noise).

Building on this understanding, consider a binary linear classifier $F(x) = \text{sign}(\Psi(x))$, where $\Psi(x) = \langle x, z \rangle$ with $z \neq 0$, Etmann *et al.* [26] demonstrated a connection between the adversarial robustness of F and the alignment between the input image x and the layer weights z (which directly impact the saliency map $\nabla_x \Psi$). Specifically, they found that a higher alignment between the weights z and the input image x contributes to an increased adversarial robustness in the model. This result suggests that a model that exhibits adversarial robustness generates saliency maps that better align with the input image x (see Appendix for the mathematical details). Consequently, if we consider two different models demonstrating similar adversarial robustness to the same input image, they will generate saliency maps that are better aligned with one another. Thus, we expect that models sharing an architecture but trained with different adversarial attack methods yield correlated interpretability maps. Indeed, the results of Etmann *et al.* [26] suggest that even models with different architecture and trained with different adversarial strategies but with close adversarial robustness will still yield correlated interpretability maps.

Finally, our work opens several new directions for promoting more robust DL models. First, while we do not explore the impact of adversarial training on vision transformers [3], we have no doubts that our approach can be generalized to the context of vision transformers. This generalization, however, requires the application of attribution formulations that are specific to transformers [55]. Second, there is a clear need to benchmark the new input attribution methods that are being proposed. While not a complete benchmarking approach, our study provides a roadmap for the benchmarking of attribution methods based on evidence that the model is learning the justifiably important features.

The STL-10 dataset is available at [56], the CIFAR-10 dataset is openly available, and the RIVAL10 is available at [57]. Images used to investigate the foreground and background of this study are available at [58].

The code to reproduce the results and figures in this study is available at [58].

ACKNOWLEDGMENTS

This work has been partially supported by CAPES and CNPq (Brazilian research agencies).

F.A.O.S, C.Z., and L.A.N.A. conceived and designed the study. F.A.O.S and C.Z. performed the numerical simulations. F.A.O.S, C.Z., W.L., and L.A.N.A. performed the data analysis. F.A.O.S, C.Z., W.L., and L.A.N.A. created the figures. F.A.O.S, C.Z., and L.A.N.A. wrote the first draft of the paper. F.A.O.S, C.Z., W.L., and L.A.N.A. wrote, read, and approved the final version of the paper.

APPENDIX

This Appendix presents the complete results of the experiments and the results of all evaluated attribution methods. In addition, we also present the training hyperparameters and standard evaluation of each model.

1. Models setup

a. Training parameters

Table II presents all the hyperparameter values and the learning rate schedule. We have chosen the values of the hyperparameters based on [59]. Furthermore, we have optimized the tradeoff between keeping the hyperparameters with equal value within each dataset scenario as much as possible and producing a model with competitive accuracy. Therefore, the batch size and the number of epochs are equal.

b. Standard evaluation

The columns Test Acc and FGSM from Table II present the results obtained from each model in the datasets CIFAR-10, STL-10, and RIVAL10. As expected, the classically trained models have better results than the adversarially trained models in the standard evaluation [25]. However, the adversarially trained models are more robust to adversarial attacks.

In order to enhance our conclusions about the correlation analysis, we extend it to two additional architectures, namely ResNet-34 and GoogleNet. Figure 6 displays the correlation matrices for the ResNet-34 and GoogleNet architectures. The results demonstrate that the correlation scores are consistently higher among adversarial robust models compared to nonadversarial ones, thus confirming the alignment with our previously reported findings in the paper.

The layer-by-layer analysis revealed that adversarial training has a significant impact on the first layer of the PreActResnet-18 model. To further validate this observation, we conducted additional experiments on the GoogleNet and ResNet-34 architectures and presented the results in Fig. 7 in this Appendix. The findings indicate that adversarial training predominantly affects the initial layers of ResNet-34 and GoogleNet, thus being coherent with the findings in the paper.

2. Analysis of the interpretability maps correlation between different architectures

In Fig. 2(b) we showed that different PGD models' architecture produces strongly correlated interpretability maps. In order to present the complete results, we extend this analysis for all training approaches (i.e., Adam, SGD, PGD, and FGSM) in Fig. 8. The complete results show that for each interpretability method that computes attribution maps with relation to the input rather than intermediate layers (i.e., all methods except Grad-CAM), the correlations between different architectures trained using the same adversarial attack method are consistently higher than the corresponding correlations between architectures trained using nonadversarial

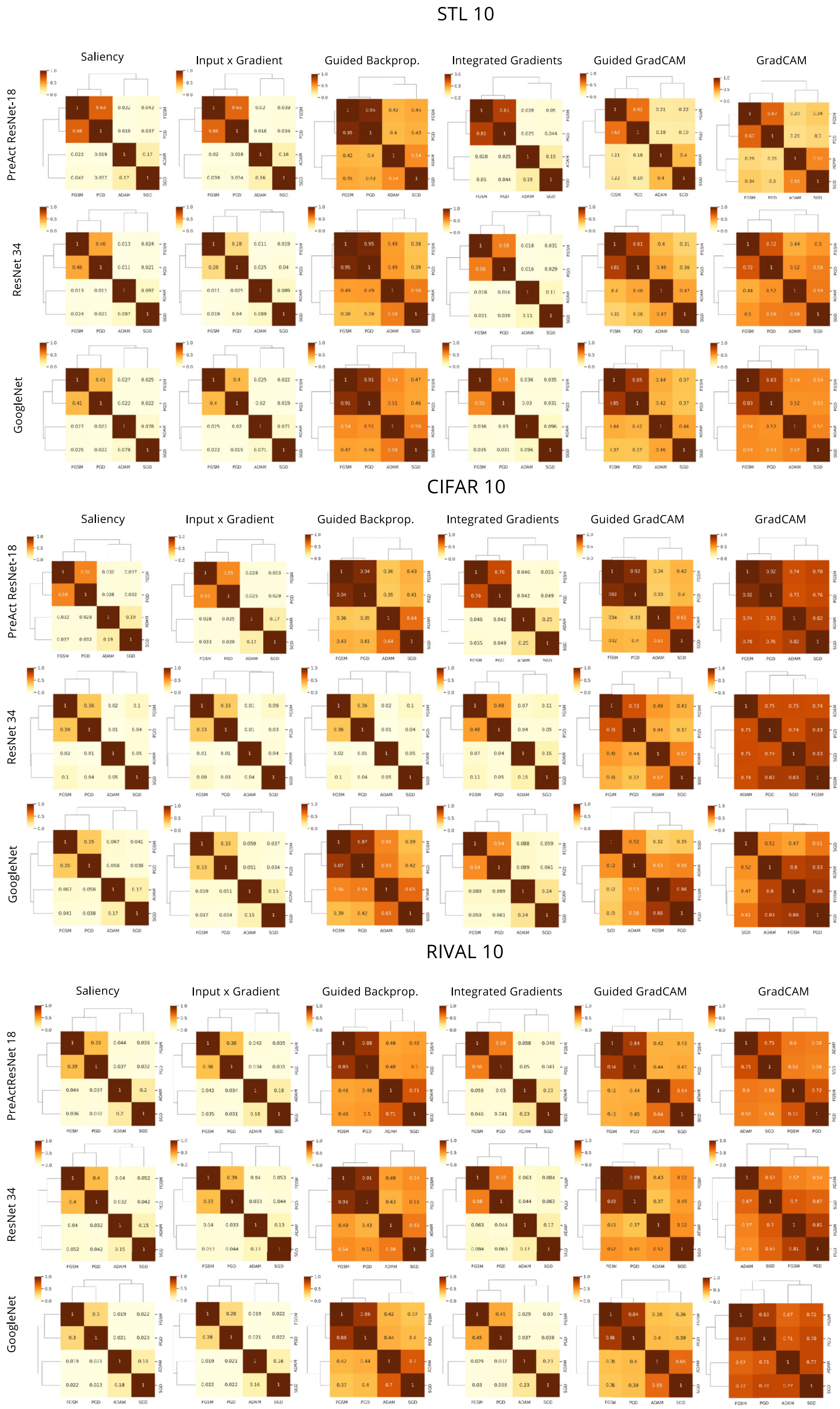

FIG. 6. Adversarial trained models have more correlated interpretability maps than non-adversarial. The plot is organized in three grids of results, the top one is for the STL-10 dataset, while the middle is for CIFAR10 and the bottom is for RIVAL-10. The findings in the paper extend to other architectures, such as ResNet-24 and Google Inception, for almost all cases.

TABLE II. Hyperparameters values to models training. Each row represents the training settings for each model. The alpha and epsilon are hyperparameters related to adversarial training, and the attack iters represent the number of iterations used in the PGD method. The cells with no value suggest that the training method does not need this parameter. The columns FGSM present the results obtained from the adversarial robustness evaluation.

Training Approach	Architecture	Dataset	Batch size	Epochs	Alpha	Epsilon	LR	Test Acc	FGSM Acc
SGD	PreAct-18	CIFAR-10	128	50			0.2	94.41	0.0
Adam	PreAct-18	CIFAR-10	128	50			0.005	92.82	0.0
FGSM	PreAct-18	CIFAR-10	128	50	0.16	0.13	0.2	85.67	45.19
PGD	PreAct-18	CIFAR-10	128	50	0.03	0.13, iters 10	0.2	83.39	50.68
SGD	GoogleNet	CIFAR-10	64	100			0.04	90.40	22.6
Adam	GoogleNet	CIFAR-10	128	100			0.0001	90.91	6.21
FGSM	GoogleNet	CIFAR-10	128	100	0.16	0.13	0.01	77.22	48.73
PGD	GoogleNet	CIFAR-10	128	100	0.03	0.13, iters 10	0.01	86.12	55.47
SGD	ResNet-34	CIFAR-10	128	100			0.05	90.41	12.36
Adam	ResNet-34	CIFAR-10	128	100			0.0015	93.02	14.30
FGSM	ResNet-34	CIFAR-10	128	100	0.16	0.13	0.05	86.28	33.60
PGD	ResNet-34	CIFAR-10	128	100	0.03	0.13, iters 10	0.0015	83.16	54.88
SGD	PreAct-18	STL-10	32	50			0.2	81.81	0.0
Adam	PreAct-18	STL-10	32	50			0.0015	82.55	0.0
FGSM	PreAct-18	STL-10	32	50	0.16	0.13	0.2	72.76	37.58
PGD	PreAct-18	STL-10	32	50	0.03	0.13, iters 10	0.2	71.30	40.79
SGD	GoogleNet	STL-10	32	100			0.005	80.05	17.91
Adam	GoogleNet	STL-10	32	100			0.0001	86.41	0.02
FGSM	GoogleNet	STL-10	32	100	0.16	0.13	0.01	75.40	38.16
PGD	GoogleNet	STL-10	32	100	0.03	0.13, iters 10	0.01	72.10	40.13
SGD	ResNet-34	STL-10	32	100			0.005	78.55	4.71
Adam	ResNet-34	STL-10	32	100			0.0015	85.74	2.85
FGSM	ResNet-34	STL-10	32	100	0.16	0.13	0.005	72.73	34.31
PGD	ResNet-34	STL-10	32	100	0.03	0.13, iters 10	0.005	80.29	37.04
SGD	PreAct-18	RIVAL-10	128	100			0.01	83.31	7.56
Adam	PreAct-18	RIVAL-10	128	100			0.0001	86.80	8.97
FGSM	PreAct-18	RIVAL-10	128	100	0.16	0.13	0.01	74.14	38.81
PGD	PreAct-18	RIVAL-10	128	100	0.03	0.13, iters 10	0.01	77.53	42.04
SGD	GoogleNet	RIVAL-10	128	200			0.01	84.54	7.08
Adam	GoogleNet	RIVAL-10	128	200			0.0015	87.42	6.41
FGSM	GoogleNet	RIVAL-10	128	100	0.16	0.13	0.015	72.80	37.70
PGD	GoogleNet	RIVAL-10	128	100	0.03	0.13, iters 10	0.01	76.88	43.27
SGD	ResNet-34	RIVAL-10	128	200			0.01	82.99	10.06
Adam	ResNet-34	RIVAL-10	128	200			0.0001	88.42	11.93
FGSM	ResNet-34	RIVAL-10	128	100	0.16	0.13	0.02	74.57	37.59
PGD	ResNet-34	RIVAL-10	128	100	0.03	0.13, iters 10	0.01	74.54	37.91

methods (i.e., Adam or SGD). This finding is coherent with the present in the paper on Fig. 2(b).

3. On why adversarial robust models produce correlated interpretability maps

To elucidate the implications of our experimental findings and provide a more nuanced understanding, we build here on the work by Etmann *et al.* [26] regarding the robustness of a binary linear classification model. Our analysis enables us to demonstrate a connection between the saliency map obtained from two different models, which are both adversarially robust to the same input vector x. For concreteness, we focus on saliency map interpretability.

We start by recalling Etmann *et al.*'s definitions. [26]

Definition 1. Let $F : X \to C$ with C finite, be a classifier over the normed vector space $(X, \|\cdot\|)$. We call

$$\rho(x) = \inf_{e \in X}\{\|e\| : F(x+e) \neq F(x)\}, \tag{A1}$$

the (adversarial) robustness of F for input x, and call $E_{x\sim D}[\rho(x)]$ the (adversarial) robustness of F over the distribution D.

Definition 2. Let the binary classifier $F : X \to \{-1, 1\}$ be defined almost everywhere (a.e.) by

$$F(x) = \mathrm{sign}(\Psi(\mathrm{x})) = \mathrm{sign}(\langle \mathrm{x}, \mathrm{z}\rangle)\,, \tag{A2}$$

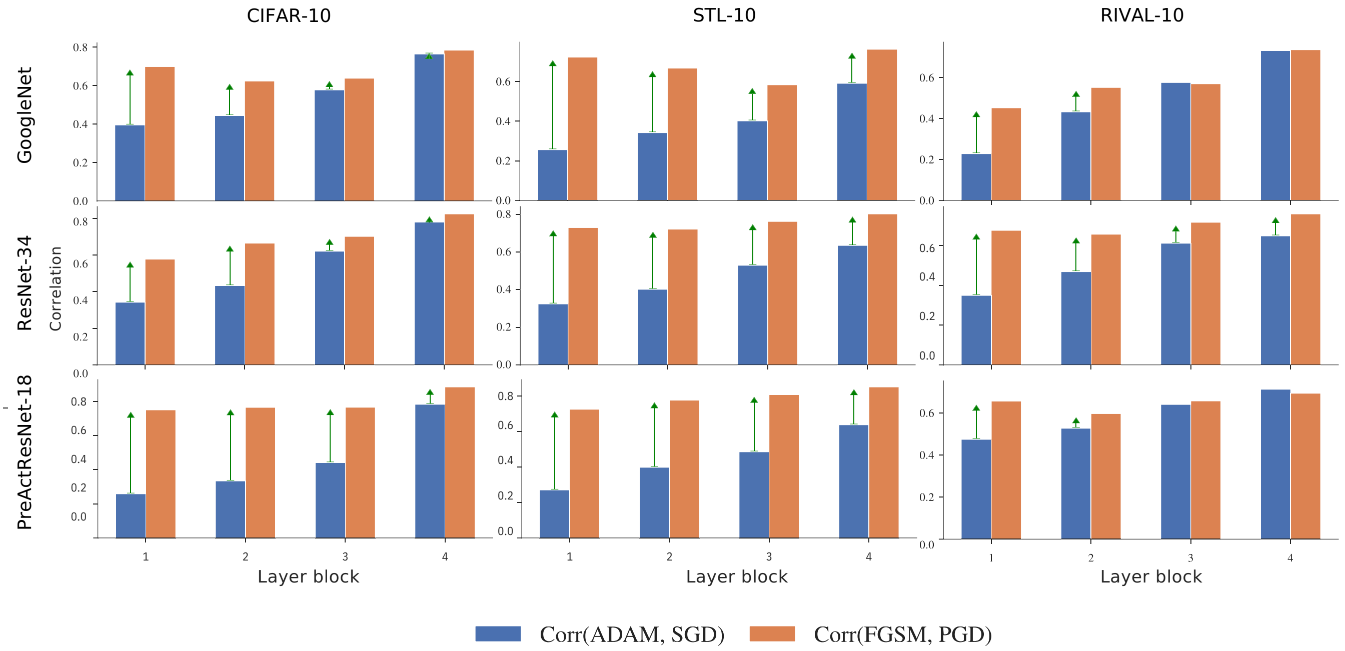

FIG. 7. Layer-by-layer analysis. This plot illustrates the average correlation between the integrated gradient interpretability maps of two model pairs, namely (Adam, SGD) and (FGSM, PGD), across different model layers. Each row presents the results for each architecture, GoogleNet, ResNet-34, and PreActResNet-18, respectively. The green arrow means the difference between the average correlation of (Adam, SGD) and (FGSM, PGD).

where $\Psi : X \rightarrow \mathbb{R}$ is differentiable in x. We call $\nabla\Psi$ the saliency map of F with respect to Ψ in x, and call

$$\alpha(x) = \frac{|\langle x, \nabla_x \Psi(x)\rangle|}{\|\nabla_x \Psi(x)\|}, \tag{A3}$$

the alignment with respect to Ψ in x.

Etmann *et al.* [26] demonstrated that

$$\rho(x) = \frac{|\langle x, z\rangle|}{\|z\|} = \frac{|\langle x, \nabla_x \Psi\rangle|}{\|\nabla_x \Psi\|} = \|x\| \cdot |\cos(\delta)|\,, \tag{A4}$$

where δ is the angle between the vectors x and $\nabla_x \Psi$. These results show that the adversarial robustness in an input vector x depends on the alignment between the linear model weights z (which directly impact the saliency map $\nabla_x \Psi$) and the input vector x.

From these results, we can easily derive a relationship between the saliency maps of two linear binary models. Consider two binary linear models

$$F^1 = \text{sign}(\Psi^1(\text{x})) \quad \text{and} \quad \text{F}^2 = \text{sign}(\Psi^2(\text{x})). \tag{A5}$$

If both have robustness in regard to an input vector x of approximately $\rho(x)$, then it follows from Eq. (A4) that

$$\rho(x) = \|x\| \cdot \frac{|\langle x, \nabla_x \Psi^1\rangle|}{\|x\|\|\nabla_x \Psi^1\|} = \|x\| \cdot \frac{|\langle x, \nabla_x \Psi^2\rangle|}{\|x\|\|\nabla_x \Psi^2\|} + \epsilon\,, \tag{A6}$$

and

$$\frac{|\langle x, \nabla_x \Psi^1\rangle|}{\|\nabla_x \Psi^1\|} - \frac{|\langle x, \nabla_x \Psi^2\rangle|}{\|\nabla_x \Psi^2\|} = \epsilon \ll 1\,. \tag{A7}$$

Both models must also fulfill the condition that the angle δ between the vector x and their saliency maps must be in the interval $[0^\circ, 90^\circ)$. It follows that $\langle x, \nabla_x \Psi^1\rangle > 0$ and $\langle x, \nabla_x \Psi^2\rangle > 0$. This enables us to pull the x out of the correlation calculation

$$x \cdot \left(\frac{\nabla_x \Psi^1}{\|\nabla_x \Psi^1\|} - \frac{\nabla_x \Psi^2}{\|\nabla_x \Psi^2\|} \right) = \epsilon \ll 1\,. \tag{A8}$$

For this equation to hold for all vectors x, then the term inside the parentheses must be close to zero, and the two salience maps must be approximately the same:

$$\frac{\nabla_x \Psi^1}{\|\nabla_x \Psi^1\|} \approx \frac{\nabla_x \Psi^2}{\|\nabla_x \Psi^2\|}\,. \tag{A9}$$

This argument shows that two models with similar adversarial robustness will produce similar salience maps. Connecting this result to the finding that different architectures trained with either the PGD or the FGSM approaches have similar adversarial robustness (Table II) allows us to understand why a continuum of similarities of adversarial robustness produces a continuum of degrees of correlations of attribution maps. The question remains, though, of why different models trained using classical methods do not display strong correlations across interpretability methods. The results above suggest that this may be due to the possibility that the weights for two different classically trained models are not as well aligned as the weights for two different adversarially trained models. That is, classically trained models attain similar performance with very different sets of weights, whereas adversarially trained models converge to more similar sets of weights. While this is an important insight that deserves greater attention, it is beyond the scope of the current manuscript.

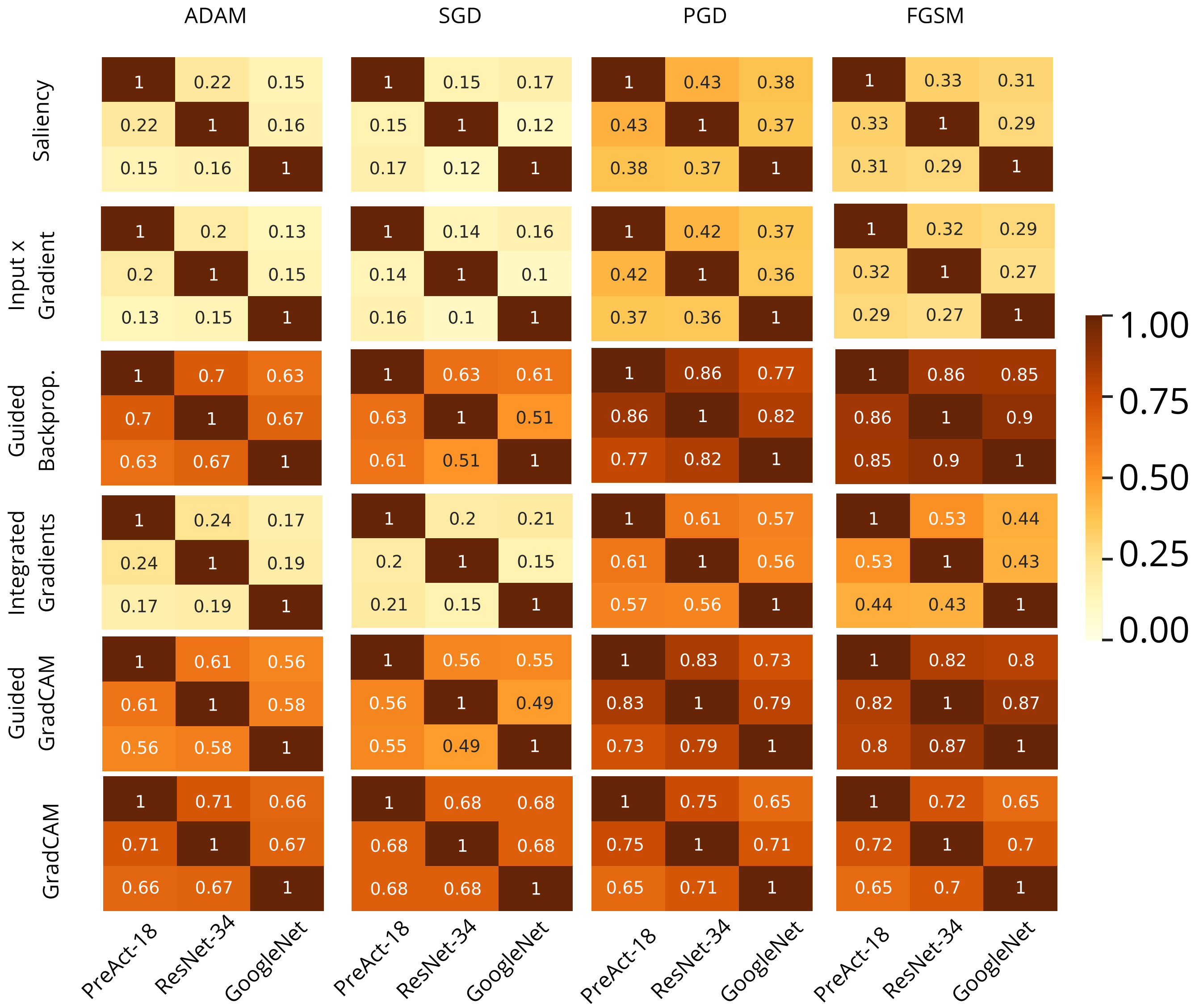

FIG. 8. Interpretability correlation computed between different architectures trained with the same strategy. Each row of the plot presents the results for a specific interpretability method (from saliency to Grad-CAM), while each column presents the results for each training method (from Adam to FGSM).

[1] P. Li *et al.*, Trimnet: learning molecular representation from triplet messages for biomedicine, Briefings Bioinf. **22**, bbaa266 (2021).

[2] D. S. Park *et al.*, Improved noisy student training for automatic speech recognition, in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association*, edited by H. Meng, B. Xu, and T. F. Zheng (ISCA, 2020), pp. 2817–2821.

[3] A. Dosovitskiy *et al.*, An image is worth 16x16 words: Transformers for image recognition at scale, in *9th International Conference on Learning Representations* (OpenReview.net, 2021).

[4] Y. Zhu, L. Pang, Y. Lan, H. Shen, and X. Cheng, Adaptive information seeking for open-domain question answering, in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, edited by M. Moens, X. Huang, L. Specia, and S. W. Yih (Association for Computational Linguistics, 2021), pp. 3615–3626.

[5] S. Takase, and S. Kiyono, Rethinking perturbations in encoder-decoders for fast training, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, edited by K. Toutanova *et al.* (Association for Computational Linguistics, 2021), pp. 5767–5780.

[6] Y. Pan, T. Yao, Y. Li, and T. Mei, X-linear attention networks for image captioning, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020* (Computer Vision Foundation / IEEE, 2020), pp. 10968–10977.

[7] F. Fuchs, Y. Song, E. Kaufmann, D. Scaramuzza, and P. Dürr, Super-human performance in gran turismo sport using deep

reinforcement learning, IEEE Robot. Autom. Lett. **6**, 4257 (2021).

[8] Google Search Will Be Your Next Brain, https://www.wired.com/2015/01/google-search-will-be-your-next-brain/.

[9] Amazon scientists applying deep neural networks to custom skills, https://www.amazon.science/blog/amazon-scientists-applying-deep-neural-networks-to-custom-skills.

[10] Powered by AI: Instagram's Explore recommender system, https://ai.facebook.com/blog/powered-by-ai-instagrams-explore-recommender-system/.

[11] J. Stilgoe, Self-driving cars will take a while to get right, Nat. Mach. Intell. **1**, 202 (2019).

[12] M. Roberts *et al.*, Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, Nat. Mach. Intell. **3**, 199 (2021).

[13] R. Geirhos *et al.*, Shortcut learning in deep neural networks, Nat. Mach. Intell. **2**, 665 (2020).

[14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, Intriguing properties of neural networks, in *Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada*, edited by Y. Bengio and Y. LeCun (2014).

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA*, edited by Y. Bengio and Y. LeCun (2015).

[16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, Towards deep learning models resistant to adversarial attacks, in *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada* (2018).

[17] A. Nguyen, J. Yosinski, and J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, *2015 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR, 2015), pp. 427–436.

[18] M. A. Alcorn *et al.*, Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR, 2019).

[19] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, Natural adversarial examples, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2021), pp. 15262–15271.

[20] S. Beery, G. Van Horn and P. Perona, Recognition in terra incognita, in *Proceedings of the European conference on computer vision* (ECCV, 2018), pp. 456–473.

[21] M. Simon, E. Rodner, and J. Denzler, Part detector discovery in deep convolutional neural networks, *Asian Conference on Computer Vision* (Springer, 2014), pp. 162–177.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. **128**, 336 (2020).

[23] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, Striving for simplicity: The all convolutional net, in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA*, edited by Y. Bengio and Y. LeCun (2015).

[24] T. Zhang, and Z. Zhu, Interpreting adversarially trained convolutional neural networks, in *Proceedings of the 36th International Conference on Machine Learning*, edited by K. Chaudhuri, and R. Salakhutdinov (PMLR, 2019), pp. 7502–7511.

[25] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, Robustness may be at odds with accuracy, arXiv:1805.12152.

[26] C. Etmann, S. Lunz, P. Maass, and C. Schoenlieb, On the connection between adversarial robustness and saliency map interpretability, *International Conference on Machine Learning* (PMLR, 2019), pp. 1823–1832.

[27] A. Ignatiev, N. Narodytska, and J. Marques-Silva, On relating explanations and adversarial examples, in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada*, edited by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett (2019), pp. 15857–15867.

[28] A. Chan, Y. Tay, Y. S. Ong, and J. Fu, Jacobian adversarially regularized networks for robustness, in *Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia* (2020).

[29] J. Chen, X. Wu, V. Rastogi, Y. Liang, and S. Jha, Robust attribution regularization, in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada*, edited by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett (2019), pp. 14300–14310.

[30] S. Yin, K. Yao, S. Shi, Y. Du, and Z. Xiao, AGAIN: Adversarial training with attribution span enlargement and hybrid feature fusion, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada* (IEEE, 2023), pp. 20544–20553.

[31] A. Noack, I. Ahern, D. Dou, and B. Li, An empirical study on the relation between network interpretability and adversarial robustness, SN Comput. Sci. **2**, 1 (2021).

[32] P. Mangla, V. Singh, and V. N. Balasubramanian, On saliency maps and adversarial robustness, *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020* (Springer, 2021), pp. 272–288.

[33] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA* (IEEE, 2016), pp. 770–778.

[34] K. He, X. Zhang, S. Ren, and J. Sun, Identity mappings in deep residual networks, *Computer Vision–ECCV 2016: 14th European Conference* (Springer, 2016), pp. 630–645.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA* (IEEE, 2015), pp. 1–9.

[36] D. P. Kingma, and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.

[37] H. Robbins, and S. Monro, A stochastic approximation method, Ann. Math. Stat. **22**, 400 (1951).

[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE **86**, 2278 (1998).

[39] M. Sundararajan, A. Taly, and Q. Yan, Axiomatic attribution for deep networks, in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, edited by D. Precup, and Y. W. Teh (PMLR, 2017), pp. 3319–3328.

[40] A. Shrikumar, P. Greenside, and A. Kundaje, Learning important features through propagating activation differences, *International Conference on Machine Learning* (PMLR, 2017), pp. 3145–3153.

[41] A. Coates, A. Ng, and H. Lee, An analysis of single-layer networks in unsupervised feature learning, in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (JMLR Workshop and Conference Proceedings, 2011), pp. 215–223.

[42] A. Krizhevsky, G. Hinton *et al.*, Learning multiple layers of features from tiny images (2009).

[43] M. Moayeri, P. Pope, Y. Balaji, and S. Feizi, A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA* (IEEE, 2022), pp. 19065–19075.

[44] P. Sturmfels, S. Lundberg, and S.-I. Lee, Visualizing the impact of feature attribution baselines, Distill (2020), https://distill.pub/2020/attribution-baselines.

[45] M. Moayeri, K. Banihashem, and S. Feizi, Explicit tradeoffs between adversarial and natural distributional robustness, Advances in Neural Information Processing Systems **35**, 38761 (2022).

[46] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, Right for the right reasons: Training differentiable models by constraining their explanations, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia*, edited by C. Sierra (ijcai.org, 2017), pp. 2662–2670.

[47] P. Schramowski *et al.*, Making deep neural networks right for the right scientific reasons by interacting with their explanations, Nat. Mach. Intell. **2**, 476 (2020).

[48] L. Rieger, C. Singh, W. J. Murdoch, and B. Yu, Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge, in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020* (PMLR, 2020), pp. 8116–8126.

[49] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne, Learning how to explain neural networks: PatternNet and PatternAttribution, in *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada* (2018).

[50] A. S. Ross, and F. Doshi-Velez, Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients, in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, edited by S. A. McIlraith, and K. Q. Weinberger (AAAI Press, 2018), pp. 1660–1669.

[51] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, Ai for radiographic covid-19 detection selects shortcuts over signal, Nat. Mach. Intell. **3**, 610 (2021).

[52] G. Gigerenzer, and R. Selten, *Bounded rationality: The adaptive toolbox* (MIT Press, 2002).

[53] D. Kahneman, *Thinking, Fast and Slow* (Macmillan, 2011).

[54] J. L. Borges, *Funes, the Memorious* (La Nación, 1942).

[55] H. Chefer, S. Gur, and L. Wolf, Transformer interpretability beyond attention visualization, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2021, Nashville, TN, USA* (IEEE, 2021), pp. 782–791.

[56] https://cs.stanford.edu/∼acoates/stl10/.

[57] https://mmoayeri.github.io/RIVAL10/index.html.

[58] https://github.com/faos1993/advatk-interpretability.

[59] E. Wong, L. Rice, and J. Z. Kolter, Fast is better than free: Revisiting adversarial training, in *Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia* (2020).