Exploring Image Classification Robustness and Interpretability with Right for the Right Reasons Data Augmentation

Flávio Arthur Oliveira Santos and Cleber Zanchettin
Centro de Informática, Universidade Federal de Pernambuco,
Cidade Universitária, Recife - PE, Brazil, 50740-560
{faos,cz}@cin.ufpe.br

Abstract

Right for the right reasons (RRR) methods have been proposed to mitigate the issues of shortcut learning in deep learning models. During training, these methods guide the models to learn patterns from signal information while ignoring noisy features. This work investigates the robustness of image classification models to background sensitivity, referring to a model's capability to accurately classify an image without leveraging the shortcut learning between the image background and the assigned input label. We propose a new approach, the Right for the Right Reasons Data Augmentation (RRDA). This approach augments the image foreground context with the context extracted from different images, thereby stimulating the model to focus on signal features rather than the context. Our experiments demonstrate that RRDA can significantly improve the robustness of image classification models, outperforming other RRR methods, such as GradMask and ActDiff. We also evaluate the impact of architectural choice on robustness, showing that ViT is more robust than ResNet in handling background sensitivity. Finally, we perform an interpretability analysis to understand how models assign importance to signal and context features during the inference process. This involves computing the signal-to-noise ratio as the importance of the signal divided by the importance of the context. Contrary to our expectations, our findings suggest that a high signal-to-noise ratio does not necessarily imply robustness. However, they indicate that applying RRDA can help the models learn to focus on signal features, leading to more interpretable and robust models.

1. Introduction

Deep learning (DL) models have become state-of-the-art in tasks such as natural language [10] and image understanding [26]. However, the acceptance of these models in high-stakes domain applications (e.g., healthcare and legal systems) has been limited due to their lack of interpretability and their bias towards spurious signals [7].

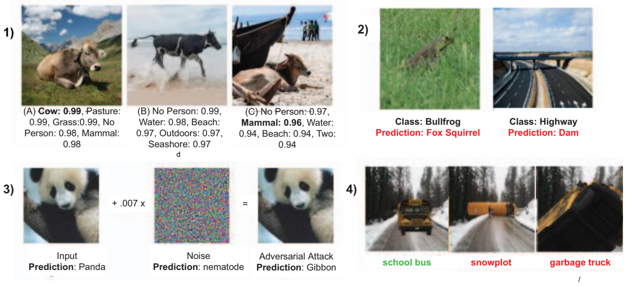
1)
(A) **Cow: 0.99**, Pasture: 0.99, Grass:0.99, No Person: 0.98, Mammal: 0.98
(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97
d
(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94
2)
Class: Bullfrog
Prediction: Fox Squirrel
Class: Highway
Prediction: Dam
3)
Input
Prediction: Panda
+ .007 x
Noise
Prediction: nematode
=
Adversarial Attack
Prediction: Gibbon
4)
school bus
snowplot
garbage truck

Figure 1. **Example of image classification failures.** Debugging DL models is important to diagnose model failures and help understand model decisions. Several works explore deep learning model decisions with different types of input information. For example, situation 1) shows that the model fails to classify a *cow* when it is present on a background different than usual. In situation 2), a bullfrog is misclassified as a fox squirrel and a highway as a dam. Unlike the first two examples, situation 3) presents an example of an adversarial attack, demonstrating that after adding noise to the input, the model fails drastically, even though it made the correct decision on the original image. Situation 4) shows that the model fails to classify an image correctly while maintaining the same background but changing the object position. The main figures of the plots were obtained from [14, 1, 9].

Investigations to understand the decisions of DL models have uncovered several situations in which DL models can fail. For instance, Szegedy et al. (2013) [23] found counterintuitive properties of DL models, demonstrating that the addition of minimal noise to the model input can lead the model to change its decision to an incorrect prediction. This failure is known as an Adversarial Attack. While this exposes potential issues with the robustness of DL models, existing literature counters this criticism by arguing that the data used in these attacks is artificially generated and falls out-of-distribution. Despite these counterarguments, several works demonstrate that DL models can fail drastically even when dealing with natural images (Figure 1).

These cases suggest that DL models may make decisions

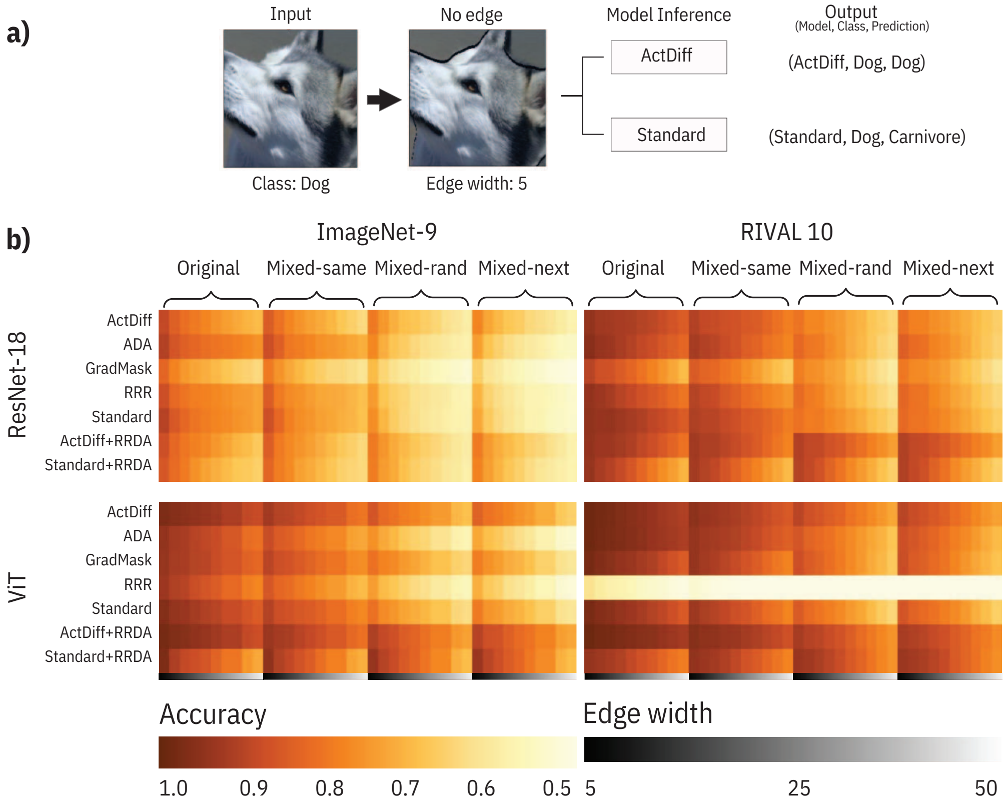

Figure 2. **a) Pipeline of Edge Analysis.** Given an input image of class y, we first conduct an edge analysis that erases the image edge of width W. We then compute the model's inference to produce a triplet consisting of the model name, class, and predicted class. This edge analysis is computed for all seven models, using an edge width ranging from 5 to 50 pixels. **b) Results Obtained from Edge Dependence Analysis.** The results are grouped by dataset, namely ImageNet9 and Rival10. For each dataset, each column represents a different challenge arranged in a sequence of increasing difficulty, starting with the original data and ending with the original data whose background is from the next class. Within each column, each cell represents the accuracy obtained for a specific edge size, starting from 5 and ending at 50.

based on incorrect information, such as background information or spurious correlations between contextual features and the input label. As such, we need robust models that make inferences based on the correct information, such as signal features or information relevant to the problem being solved. Several recent works have proposed new loss functions to guide the model to focus on signal features [17, 18, 24, 21, 16, 6], thereby using signal information instead of contextual information in the inference process. These methods are referred to in the literature as Right for the Right Reasons (RRR). These loss functions generally use second-order gradient optimization [5] and incorporate a right reasons factor into the loss function. The right reasons factor encourage the model to use the signal information in decision-making. Equation 1 presents a generic loss function for RRR training. This generic loss function consists of a $Right\_answer$ factor to guide the model to decide correctly and a $Right\_reason$ factor to instruct the model to focus on signal information. The terms λ_1 and λ_2 are parameters to weigh the contributions of both factors, where $\lambda_1 + \lambda_2$ does not necessarily sum up to 1.

$$\begin{aligned} L(\theta, X, y, A) = \lambda_1 Right\_answer(\theta, X, y) \\ +\lambda_2 Right\_reason(\theta, X, y, A). \end{aligned} \tag{1}$$

Though RRR has shown promising results, these methods have not been evaluated on a large-scale benchmark and generally use interpretability methods in the $Right\_reason$ term, making them vulnerable to interpretability fairness. In this work, we take a data-centric approach, and instead of changing how the model learns, we alter the data it learns from. We propose the Right Reasons Data Augmentation concept, which aims to augment input data context information with real, class-different context information. We hypothesize that if the model learns to classify the same signal with different contextual information, it will focus on the signal rather than the context and will be invariant to contextual features instead of deciding based on them.

2. Related works

2.1. Interpretability

Velez and Kim define interpretability in machine learning systems as the ability to explain or present in understandable terms to a human [3]. However, due to the high-dimensional nature and large number of composite layers in deep learning, interpreting its decisions can be challenging. Several methods have been proposed to mitigate this issue. One of the first methods proposed to obtain model interpretability was Saliency [20]. Given an input vector x, model f, and a class of interest c, it computes interpretability (also known as attribution maps) through the gradient of the class output with respect to the input vector (i.e., $\nabla f_c(x)$). GradCam [19] differs from Saliency by computing interpretability maps through the partial derivatives of the model output with respect to specific layer feature maps ($\frac{\partial y^c}{A^k}$). It then computes each feature map importance α_k by summing its derivatives and computing a general weight sum using α_k as the weight for feature map k. Afterward, GradCam applies the ReLU [13] function on the weight sum result to produce the model interpretability. Although Saliency and GradCam strive to produce model interpretability maps, they have received criticism for their formalism. To mitigate the lack of formalism from interpretability methods, Sundararajan et al. (2017) [22] proposed the Integrated Gradients (IG) method using an axiomatic approach. Given an input vector x, a baseline vector $x^{'}$, and a model f, IG computes f decision interpretability by cumulating the gradient of all points on the straight line between x and $x^{'}$.

2.2. Right for the Right Reasons Approach

Deep learning models can identify patterns even when dataset labels are shuffled [27]. These models typically have many layers and employ non-linear activation functions, making the interpretability of their decision-making process hard. The interpretability methods discussed so far have demonstrated how to acquire model interpretability and highlight important and non-important features used during model inference. After obtaining the model's interpretability, we can verify whether the model uses the same features that a domain specialist would. This verification could reveal a scenario where the model makes the correct decision by relying on context or spurious features, which a domain specialist would not use. Right for the Right Reasons (RRR) methods aim to mitigate this issue by optimizing the model during its learning process to ignore non-important features, thereby focusing on signal features.

As far as we know, Right for the Right Reasons [17] is the first method proposing a unique loss function that compels the model to ignore non-important features in the input vector, thus making the model RRR. Equations 2-3 present the RRR optimization loss. Given an input vector x, target class c, and a binary mask indicating the non-important (1) and important features (0), the RRR loss function computes the standard $crossentropy$ and adds a right reasons factor (RRF) to penalize the importance of non-important features. The RRF is created by the weighted sum between binary mask A and the partial derivatives of log output with respect to input vector x. As A only has 1 in non-important features, the model will learn to assign less importance to these features.

$$L_c = -\sum_{c=1}^{C} y_c log(\hat{y}_c); \tag{2}$$

$$L_{rrr}(\theta, X, y, A) = L_c + \lambda_1 (A_d \cdot \frac{\partial}{\partial x_d} \sum_{c=1}^{C} log(\hat{y}_c))^2 + \lambda_2 \sum_i \theta_i^2. \tag{3}$$

While RRR compels each output category to use less information on non-important features, it does not consider the relationship between different categories' outputs. GradMask [21], presented in equation 4, proposes a distinct loss function for the model to learn the difference between the outputs of two distinct categories due to important features rather than non-important ones. To learn this, its RRF computes the partial derivatives of the difference between two categories' output concerning input vector x, multiplies it by A segmentation mask, and thus adds the non-important feature importance as well as a loss signal.

$$L = L_c + \| \frac{\partial \|\hat{y}_1 - \hat{y}_0\|}{\partial x} \cdot A \|. \tag{4}$$

RRR and GradMask use second-order derivatives to compel the model to ignore non-important input features. This operation is costly since it requires two backward passes in the model. Furthermore, forcing the model to have a zero value in some gradient positions does not guarantee that this feature is not important. It can be a minimum, maximum, or saddle point. ActDiff [24] proposes a different approach by optimizing the model to produce the same features representation when the input has only important features and when it includes both important and non-important features, as shown in equation 5. Given an input vector x and f_l as the sub-model with the first l feature layers, ActDiff compels the model to learn the same representation for the input vector with only important features (i.e., $x \odot (1 - A)$) and the input vector with all features (i.e., x).

$$L = L_c + \lambda_{act} \| f_l(x \odot (1 - A)) - f_l(x) \|_2. \tag{5}$$

In general, the right for the right reasons method proposes a novel loss function by adding a right reasons factor to the standard loss function. However, an analysis of its right reasons factor reveals that minimizing it does not ensure that the model will be fair and robust. For instance, with ActDiff, the model may learn to produce a low-norm representation vector, making the right reasons factor small enough not to impact the loss, while in RRR and GradMask, producing low sensitivity (derivatives close to zero) does not guarantee that it does not impact the model decision.

3. Right for the Right Reasons - A Data-Centric Perspective

Right for the Right Reasons (RRR) is a property of models relating to their robustness, fairness, and reliability. RRR models are trained to extract pertinent patterns from the input signal and make inferences based on meaningful signals rather than spurious correlations. According to the Cambridge Dictionary, context is the situation within which something exists or happens, and that can help explain it[1]. Thus, context should not be the primary focus, but rather it should assist in understanding the primary focus. Issues with data can lead models to learn shortcuts from context information [7], correlating context information with the input label and resulting in an unfair model. Several works [17, 18, 21, 16, 6, 24] propose new optimization loss functions for the model to learn to ignore non-signal information. Consequently, after the training process, models will learn to extract patterns related to the signal. In this work, we take a different approach by proposing a data-centric perspective to achieve RRR. We argue that if a model is trained on Right Reasons Data (RRD), it will inherently be RRR. In the following sections, we present the concept of Right for the Right Reasons Data (RRRD) and discuss how to transform raw data into right reasons data.

RRRD assumes an input data vector $x = [x_1, x_2, \ldots, x_n]$ that comprises both class-informative (signal) and context-informative features. *We argue that if, after training a model f with a dataset D, it correlates a set of context-informative features C with label y, this is likely because D is context-biased, and the context C only appears in input samples with label y.* Therefore, D cannot be considered an RRRD dataset because its context information alone is enough for the model to discriminate between samples. Next, we present the definitions necessary to understand this concept. These definitions assume the existence of an oracle O that is robust, fair, and trustworthy.

Definition 3.1 *Given an input vector I of category c, a subset of features, denoted as IC, is defined as 'class-informative' if it is sufficient for the model O to classify I as category c.*

Definition 3.2 *Given an input vector I of category c, a subset of features, denoted as C, is defined as 'context-informative' if its intersection with IC is empty, and it is insufficient on its own for the model O to classify I as category c.*

Definitions 3.1 and 3.2 provide clarity on what we consider class-informative and context-informative features. Moreover, these definitions imply that IC and C are disjoint sets, and their union constitutes the complete input vector.

3.1. Right Reasons Data Augmentation

This section discusses the issue of models learning patterns from context rather than signal when the data correlates context and label. We propose a solution through a data augmentation method named Right Reasons Data Augmentation (RRDA). It aims to transfer context information between data samples, encouraging the model to utilize signal information for discrimination, thereby enhancing fairness and robustness. Figure 3 presents a sample of RRDA performed on a batch of images. Algorithm 1 provides a pseudo-code illustrating of its generic implementation.

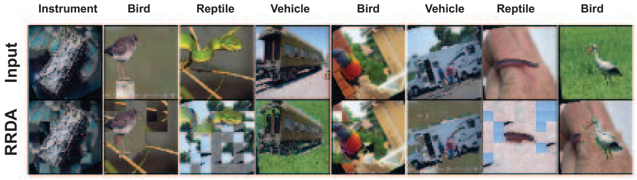

Figure 3. **Example of RRDA performing data augmentation on a batch of 8 input images.** Each column represents an input image. The first row shows the original input batch, and the second row shows the output obtained from the RRDA algorithm.

Given a batch of samples X, labels y, and a binary context information mask CI, the RRDA algorithm 1 iterates over each batch sample (Line 4) and selects a random one to replace its context information (Lines 5-7). It then adds the new samples and their respective labels to a new batch list (Lines 8-11). The primary objective of the RRDA algorithm is to embody the idea of context shift in a generic manner. For simplicity, it presumes the data is structured, and all context and class-informative features are in the same position for all samples. Therefore, when computing new_left (Line 6), we insert zeros in $X[left]$ context and add the context information from $X[right]$. This process may not apply to unstructured data (e.g., images and text) because context and signal positions often vary for each sample in the dataset. Consequently, specific implementation for each domain must address this issue.

4. Experiments and Results

Evaluating RRR is challenging as it requires a task with both signal and context information. Additionally, data manipulation and the creation of new data are necessary for as-

[1] https://dictionary.cambridge.org/us/dictionary/english/context

Algorithm 1 RRDA algorithm

```
procedure RRDA($X, y, CI$)    ▷ Compute RRDA for a batch X
    $rrda\_batch \leftarrow []$
    for $left \leq len(X)$ do
        $right \leftarrow random(len(X))$
        $new\_left \leftarrow (1 - CI[left]) \odot X[left] + X[right] \odot CI[right]$
        $new\_right \leftarrow (1 - CI[right]) \odot X[right] + X[left] \odot CI[left]$    ▷ Replace context between two samples.
        $rrda\_batch.append((new\_left, y[left]))$
        $rrda\_batch.append((new\_right, y[right]))$
    end for
    return $rrda\_batch$    ▷ The RRDA new batch and labels
end procedure
```

sessing the model's robustness in the face of context shifts, thereby extending beyond the typical test accuracy evaluation. Background sensitivity serves as a task for evaluating the impact of image backgrounds on object recognition models. We used this task to evaluate RRR methods and the proposed RRDA. If the model can ignore the background information and exhibit robustness to context shifts, it indicates a focus on the signal information. This aligns with the requirements and scope of this work.

4.1. Datasets

To evaluate background sensitivity, we need datasets with image labels and object segmentation. Therefore, we utilize the ImageNet-9 [25] challenge, which is specific to background robustness, and construct a similar background challenge with RIVAL10 [12].

ImageNet-9 (IN-9) [25] is a dataset designed for evaluating background sensitivity in object recognition. It is a subset of ImageNet [2] and consists of nine classes, each containing 5.045 training and 450 testing images. The image bounding box annotations, crucial for evaluating background sensitivity, are not abundant for each category in the original ImageNet split. Consequently, the authors of IN-9 grouped the ImageNet categories according to their ancestors in the WordNet [11] hierarchy. In addition to the raw images from ImageNet, IN-9 includes seven synthetic dataset variations intended to assess the background sensitivity of image classification models. These variations result from the processing of foreground or background elements in the original dataset. Figure 4 provides a visual example of each dataset variation.

RIVAL10 [12], a subset of ImageNet aligning with the CIFAR10 dataset classes and comprising roughly 26k images, also offers object segmentation for each image and comprehensive attribution annotation for each object. To verify the generalization of our proposal, we employ the full object segmentation from RIVAL10 to generate the mixed-same, mixed-rand, mixed-next, and only-fg variations.

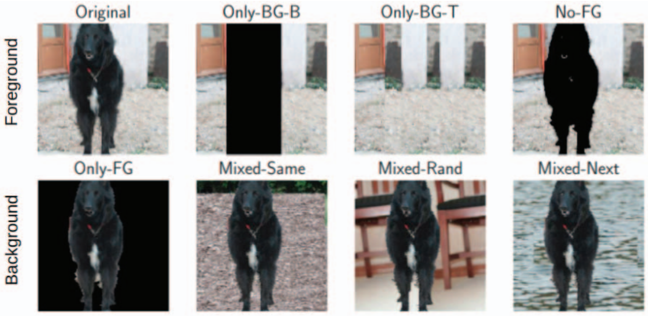

Figure 4. **ImageNet-9 challenges.** The top row displays samples of challenges that alter the foreground information, while the bottom row introduces the challenges that modify the background information. The original challenge includes images with neither foreground nor background information changes. In the Original' scenario, the original background of the image is used. BG' refers to the background, and FG' to the image foreground. In the Mixed Same' scenario, the background is swapped with the background of another image belonging to the same class. In the Mixed Rand' scenario, the background is swapped with the background of another image from a different random class. In the Mixed Next' scenario, the background is swapped with one of another image belonging to the next class, i.e., if the class index for the image is 2, then we swap backgrounds with an image from class 7.

4.2. Background Challenge Results

Table 1 presents the results from the Background challenge. The *orig.* column displays the results obtained from the original test split. All models achieve an accuracy above 90%, except for ViT with RRR, which does not generalize well on the RIVAL10 dataset. The BG-Gap column represents the difference between the mixed rand and mixed same results, indicating the variation in model accuracy when evaluated with biased (i.e., a background of the same class) and unbiased backgrounds. Therefore, a lower BG-Gap reflects a more robust model capable of handling backgrounds from different categories. It is important to highlight that the BG-Gap should be analyzed jointly with original accuracy because a perfect model (i.e., 100% on mixed rand and same) and a random model (i.e., 10% accuracy on mixed rand and same) will have BG-Gap equal to zero. The results suggest that not all RRR methods are robust to background sensitivity, as evidenced by the BG-Gap from GradMask, ADA, and RRR being worse than Standard training with the ResNet architecture on IN-9. Furthermore, the best BG-Gap on both datasets was achieved by the ViT trained with $Standard + RRDA$.

Does Robustness Depend on Dataset Characteristics? The results reveal a significant difference between the accu-

Table 1. **Challenge results.** The table organizes the results by dataset, with each row representing an evaluation. The columns 'Architecture' and 'Method' represent the architecture and training method used. The 'ImageNet-9' and 'RIVAL-10' columns represent the results for the model trained with each respective dataset. The 'Orig.', 'Mixed same', 'Mixed rand', and 'Mixed next' columns represent the accuracy results for each challenge, while the 'BG-Gap' column represents the difference between the 'Mixed rand' and 'Mixed same' results.

		ImageNet-9					RIVAL10				
Architecture	Method	Mixed same	Mixed rand	Mixed next	BG Gap	Orig.	Mixed same	Mixed rand	Mixed next	BG Gap	Orig.
ResNet-18	Standard	92.67	82.99	80.22	9.68	96.15	95.01	87.82	88.65	7.19	99.19
ResNet-18	ActDiff	90.27	84.47	83.26	5.80	93.46	94.91	86.55	87.16	8.36	98.77
ResNet-18	GradMask	86.77	76.34	73.43	10.42	90.79	90.65	83.96	84.34	6.69	96.61
ResNet-18	ADA	92.20	81.80	79.28	10.40	96.05	95.20	88.64	89.35	6.55	99.07
ResNet-18	RRR	91.90	82.12	78.77	9.78	95.31	94.82	87.89	88.67	6.92	98.90
ResNet-18	ActDiff + RRDA	89.56	85.90	84.89	3.65	92.49	96.25	94.57	94.21	1.68	98.52
ResNet-18	Standard + RRDA	88.30	83.41	82.37	4.88	90.62	95.38	93.93	93.89	1.46	96.80
ViT	Standard	94.15	86.84	84.69	7.3	98.35	95.31	87.99	88.61	7.32	99.24
ViT	ActDiff	95.98	90.27	89.46	5.7	98.99	96.92	92.26	91.47	4.65	99.62
ViT	GradMask	93.38	86.52	84.77	6.7	97.04	96.52	90.81	91.09	5.71	99.49
ViT	ADA	91.73	81.98	80.12	9.7	97.24	96.27	88.84	90.09	7.45	99.69
ViT	RRR	90.42	80.04	78.54	10.4	96.74	53.01	34.09	35.19	18.94	64.76
ViT	Standard + RRDA	**97.28**	**96.00**	**95.88**	**1.28**	**99.06**	96.67	**96.44**	**96.48**	**0.23**	97.81
ViT	ActDiff + RRDA	96.12	93.26	93.04	2.86	98.79	**97.69**	94.45	94.09	3.24	**99.75**

racies on IN-9 and RIVAL, suggesting that specific dataset features, such as categories, number of classes, image distribution, and the relationship between signal and background, may considerably impact model robustness.

On the connection between the challenges. Figure S2 present the correlation between the results of the challenges for each dataset. These results indicate a strong correlation between the Mixed same and Original, as well as between the Mixed rand and Mixed next scenarios. This result suggests how the models correlate signal with background information because mixed same have background information from the same categories, and mixed next as well mixed rand have the background from different classes.

4.3. Analysis of BG-Gap distributions

Is background robustness architecture dependent? Supervised learning design encompasses three major components: the model, the data, and the optimization loss. Thus far, we have primarily discussed different data and optimization loss functions to guide the model to adhere to the Right for the Right Reasons (RRR) principle. We aim to evaluate the impact of the architectural choice on robustness. We specifically highlight the difference between the results of the ResNet and ViT architectures, as shown in Figure 5a. The figure illustrates that the ViT architecture is more robust than ResNet, achieving a background gap minimum that is at least twice as low as that of ResNet on both datasets. Furthermore, both the maximum and median background gaps of ViT are lower than those of ResNet [15]. These findings are in line with existing literature, where authors have argued that ViT exhibits greater robustness than ResNet in terms of image transformations [15].

Does RRDA impact background robustness? Figure 5b compares the BG-Gap distributions with and without RRDA. It demonstrates a substantial impact of RRDA on BG-Gap, with high-density values for low BG-Gap approaching 0. In contrast, the BG-Gap for models without RRDA is close to 10 for both datasets. Additionally, the median values exhibit stark differences between the two situations. These results indicate that architecture design plays a crucial role in model fairness and robustness. This suggests a new direction for research, focusing on the development of "right for the right reasons" architectures rather than solely on data and optimization loss functions. Additionally, the analysis of BG-Gap distributions suggests that RRDA significantly impacts model background robustness.

Is BG-Gap dependent on original accuracy? Figure S4 presents the correlation between the BG-Gap metric and original accuracies. The correlation on RIVAL10 results is positive, while in ImageNet-9 is negative. Nevertheless, although in both cases the correlation is not strong enough, these insights could have significant implications for the training of deep learning models, suggesting that striving

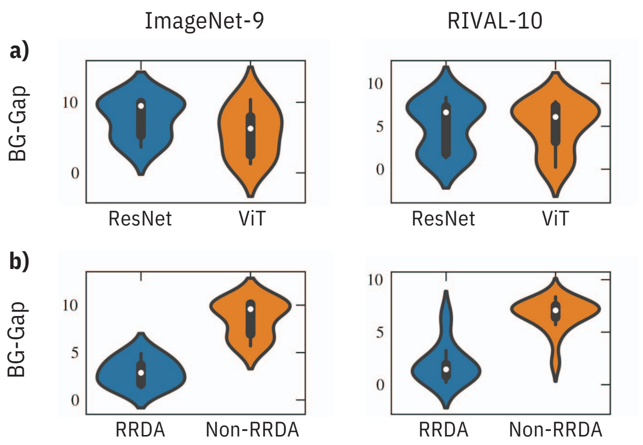

Figure 5. BG-Gap distributions for different configurations. The BG-Gap distribution is built from the BG Gap column in Table 1. The a) plot shows the comparison between ResNet and ViT architectures based on BG-Gap distribution for ImageNet-9 and RIVAL-10 datasets, while the b) plot compares the BG-Gap when we use RRDA with when we do not use (i.e., Non-RRDA).

for high accuracy might inadvertently lead to models that overfit to the background of images and the more robust model is not necessarily the best in test accuracy.

4.4. Model Dependence on Edge Information

Edge information is vital for image recognition as it represents boundaries between different pieces of information, such as the foreground and background. In this section, we question whether this information is necessary for models to make correct inferences and how robust they are to changes in edge information. We create new variations of the original IN-9 and RIVAL-10 datasets to perform the edge analysis by removing edge information with a fixed width W. As the edge represents the transition between pieces of information, we occlude parts of the signal and background, thereby eliminating this transition. Figure 2a) presents an example of the model dependence on the edge information pipeline, using an image from the original set and its version with the edge removed.

We applied edge removal to the IN-9 and RIVAL-10 original sets and all background variations, namely Mixed-same, Mixed-rand, and Mixed-next, with edge sizes varying from 5 to 50. It is important to highlight that images with high-edge sizes almost do not have information, but these scenarios are important to visualize the tendency of the results. The results of this analysis are presented in Figure 2b). These results indicate that edges are essential for all models across all challenges, as an increase in edge size corresponds to a decrease in accuracy. Another notable observation is the relationship between challenge difficulty and edge dependency. As the difficulty of the challenge increases, the models become more dependent on edge information, as indicated by lower accuracy scores. For instance, focusing on an edge size of five, accuracy decreases in line with the difficulty level of the challenge.

A significant finding is that models using the RRDA augmentation method exhibit greater robustness to edge information compared to the standard and raw RRR methods. While the standard method with RRDA maintains similar performance across all challenges, the raw standard method demonstrates greater robustness when evaluated on the original challenge. This suggests that the standard method is tailored to the original distribution and is dependent on the background. In general, ActDiff with RRDA augmentation outperformed other methods, demonstrating consistent accuracy across all challenge variations.

Table 2. **Signal Information Results.** Comparison of model performance when trained with images containing only foreground (FG) or background (BG) information.

		ImageNet-9		RIVAL10	
Arch.	Method	Only FG	Only BG	Only FG	Only BG
RN-18	Standard	85.01	32.52	89.50	41.30
RN-18	ActDiff	86.96	16.20	91.01	40.50
RN-18	GradMask	77.24	23.98	85.41	32.50
RN-18	ADA	86.72	31.78	91.01	41.32
RN-18	RRR	86.10	30.32	88.47	39.69
RN-18	ActDiff + RRDA	85.43	20.79	94.42	29.99
RN-18	Standard + RRDA	85.14	22.37	93.97	24.35
ViT	Standard	91.80	42.35	91.43	42.36
ViT	ActDiff	95.31	41.04	95.04	44.91
ViT	GradMask	90.57	33.63	92.00	45.03
ViT	ADA	87.70	36.35	93.95	47.94
ViT	RRR	86.12	33.24	37.10	27.87
ViT	Standard + RRDA	**97.30**	32.74	**96.94**	16.35
ViT	ActDiff + RRDA	95.68	44.17	95.65	48.58

4.5. Dependence of Models on Signal Information

Definitions 3.1 and 3.2 clarify what we consider as class and context informative features. In the context of an image classification task, with humans acting as a fair oracle, the object signal is sufficient for us to perform the classification. In this section, we analyze model accuracy when presented with either only object signal information or only background information. Table 2 present the results.

The results demonstrate that high test set accuracy does not necessarily translate to high accuracy when faced with only foreground information. All models trained without RRDA experience a decrease in accuracy of at least 5% (i.e., Orig. - Only FG accuracy). However, the ViT model trained with Standard + RRDA is the least affected, achieving almost the same accuracy with Only FG as with the Original test, with this difference being less than 1% on the RIVAL10 dataset.

Comparing the results of all ViT models with those of ResNet models trained with the same method, it is evident that ViT consistently achieves higher Only FG accuracy. This reinforces the claim that the choice of architecture is a fundamental building block in achieving robust models.

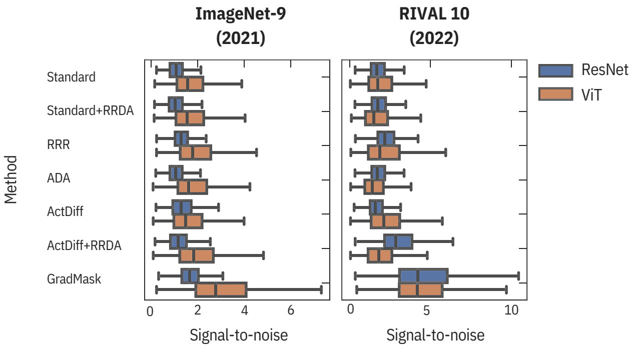

Figure 6. **Analysis of Signal-to-Noise Ratio.** For each model and dataset, we compute the signal-to-noise ratio for each image using the Saliency interpretability method. We then create a box plot to display the distribution of these ratios. The left panel presents the signal-to-noise ratio distributions for the model trained with IN-9, while the right panel illustrates the scenario with RIVAL 10.

4.6. Interpretability Methods are Fragile

Interpretability methods generate an attribution matrix, where each input dimension indicates the importance of the corresponding input feature dimension for the model's output prediction. These methods enable us to analyze the difference in feature attribution between a model robust to background changes and one that is not. To carry out this analysis, we compute the signal-to-noise ratio (i.e., the ratio between the mean importance of the signal and background) for each input image in each model and construct a box plot to analyze the differences between models. Figure 6 presents the results for the Saliency interpretability method [20].

The results show that all ViT models exhibit a higher signal-to-noise ratio than the corresponding ResNet-18 models trained with the same method on IN-9. However, this pattern does not hold for the RIVAL 10 dataset, underscoring once again that dataset characteristics are crucial in these analyses.

Does high background robustness imply high signal importance? When analyzing robust models that achieve high accuracy on only-FG and all mixed challenges, it might seem natural to expect these models to attribute high importance to the signal and low importance to the background (i.e., have a high signal-to-noise ratio). However, our analysis reveals that this assumption does not always hold. For example, even one of the less robust models, ResNet-18+GradMask, exhibits a high signal-to-noise ratio in RIVAL10. Furthermore, while RRR and GradMask have higher signal-to-noise ratios than standard+RRDA, they are less robust. This suggests that methods that learn to attribute low importance to the background (i.e., those with a high signal-to-noise ratio) are not necessarily the most robust. These counter-intuitive findings warrant further investigation, as they raise several research questions regarding the accuracy of interpretability methods and whether high signal importance is a cause or a consequence of model robustness.

5. Conclusion

This work evaluates methods such as Right for the Right Reasons (RRR), GradMask, ActDiff, and ADA for their robustness to image background sensitivity using the ImageNet-9 and RIVAL10 datasets. The results indicate that these methods struggle to create a robust model that focuses on signal information rather than context information. In response to this, we propose the Right Reasons Data Augmentation (RRDA) method to guide the training process to create robust models that prioritize signal over context information. Remarkably, our results show that RRDA improves the model performance upon the standard and ActDiff outcomes.

The vulnerability of RRR, GradMask, and ADA to background sensitivity is intriguing. To obtain deeper insights, we conducted an interpretability analysis to understand how these models attribute importance to different features. We computed the signal-to-noise ratio to quantify the relationship between signal and context importance. The results from this analysis, along with the challenges presented by the ImageNet-9 and RIVAL10 datasets, suggest that having a high signal-to-noise ratio (i.e., signal features having high importance) is not necessarily an indicator of model robustness. This helps clarify why RRR and GradMask did not improve in terms of background sensitivity. Besides, this raises questions about the fairness of interpretability methods.

References

[1] Michael A. Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. 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)*, June 2019. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[3] Finale Doshi-Velez and Been Kim. Considerations for evaluation and generalization in interpretable machine learning. In *Explainable and interpretable models in computer vision and machine learning*, pages 3–17. Springer, 2018. 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 11

[5] Harris Drucker and Yann Le Cun. Improving generalization performance using double backpropagation. *IEEE transactions on neural networks*, 3(6):991–997, 1992. 2

[6] Gabriel G. Erion, Joseph D. Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Learning explainable models using attribution priors. *CoRR*, abs/1906.10670, 2019. 2, 4

[7] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1, 4

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 11

[9] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 1

[10] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, pages 1–32, 2022. 1

[11] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 5

[12] Mazda Moayeri, Phillip Pope, Yogesh Balaji, and Soheil 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*, pages 19087–19097, 2022. 5

[13] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 3

[14] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015. 1

[15] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022. 6

[16] Laura Rieger, Chandan Singh, W. James Murdoch, and Bin 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, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8116–8126. PMLR, 2020. 2, 4

[17] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2662–2670. ijcai.org, 2017. 2, 3, 4

[18] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020. 2, 4

[19] 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*, pages 618–626, 2017. 3

[20] Marcel Simon, Erik Rodner, and Joachim Denzler. Part detector discovery in deep convolutional neural networks. In *Asian Conference on Computer Vision*, pages 162–177. Springer, 2014. 3, 8, 11

[21] Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. Gradmask: Reduce overfitting by regularizing saliency. *arXiv preprint arXiv:1904.07478*, 2019. 2, 3, 4

[22] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017. 3, 11

[23] 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. 1

[24] Joseph D. Viviano, Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. Saliency is a possible red herring when diagnosing poor generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 3, 4

[25] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 5

[26] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12104–12113, June 2022. 1

[27] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning re-

quires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 3