Attributional Robustness Training using Input-Gradient Spatial Alignment

Mayank Singh[1⋆], Nupur Kumari[1⋆], Puneet Mangla[2], Abhishek Sinha[1⋆⋆], Vineeth N Balasubramanian[2], and Balaji Krishnamurthy[1]

[1] Media and Data Science Research Lab, Adobe, India
{msingh,nupkumar}@adobe.com, abhishek.sinha94@gmail.com, kbalaji@adobe.com
[2] IIT Hyderabad, India
{cs17btech11029,vineethnb}@iith.ac.in

Abstract. Interpretability is an emerging area of research in trustworthy machine learning. Safe deployment of machine learning system mandates that the prediction and its explanation be reliable and robust. Recently, it has been shown that the explanations could be manipulated easily by adding visually imperceptible perturbations to the input while keeping the model's prediction intact. In this work, we study the problem of attributional robustness (i.e. models having robust explanations) by showing an upper bound for attributional vulnerability in terms of spatial correlation between the input image and its explanation map. We propose a training methodology that learns robust features by minimizing this upper bound using soft-margin triplet loss. Our methodology of robust attribution training (ART) achieves the new state-of-the-art attributional robustness measure by a margin of $\approx$ 6-18 % on several standard datasets, ie. SVHN, CIFAR-10 and GTSRB. We further show the utility of the proposed robust training technique (ART) in the downstream task of weakly supervised object localization by achieving the new state-of-the-art performance on CUB-200 dataset.

Keywords: Attributional robustness; Adversarial robustness; Explainable deep learning

1 Introduction

Attribution methods [9, 45, 51, 48, 47, 54, 46] are an increasingly popular class of explanation techniques that aim to highlight relevant input features responsible for model's prediction. These techniques are extensively used with deep learning models in risk-sensitive and safety-critical applications such as healthcare [4, 32, 56, 24], where they provide a human user with visual validation of the features used by the model for predictions. In computer-assisted diagnosis, [56] showed that predictions with attribution maps increased accuracy of retina specialists

⋆ Equal contribution
⋆⋆ Work done at Adobe

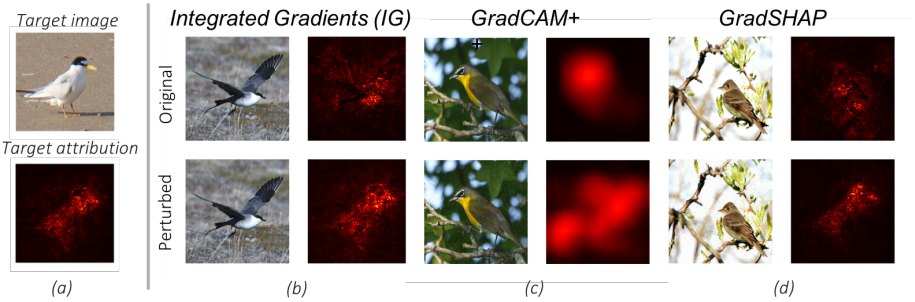


Fig. 1: Illustration of targeted manipulation [12] of attribution maps on CUB-200 [61] using the target attribution of (a). Here, (b) Integrated Gradients [54], (c) GradCAM++ [9] and (d) GradSHAP [29] blocks show : Top (b), (c), (d) original image and its attribution map; Bottom (b), (c), (d) perturbed image and its attribution map.

above that of unassisted reader or model alone. In [24], the authors improve the analysis of skin lesions by leveraging explanation maps of prediction.

It has been recently demonstrated that one could construct targeted [12] and un-targeted perturbations [16, 10] that can arbitrarily manipulate attribution maps without affecting the model's prediction. This issue further weakens the cause of safe application of machine learning algorithms. We show an illustrative example of attribution-based attacks for image classifiers over different attribution methods in Fig. 1. This vulnerability leads to newer challenges for attribution methods, as well as robust training techniques. The intuition of attributional robustness is that if the inputs are visually indistinguishable with the same model prediction, then interpretation maps should also remain the same.

As one of the first efforts, [10] recently proposed a training methodology that aims to obtain models having robust integrated gradient [54] attributions. In addition to being an early effort, the instability of this training methodology, as discussed in [10], limits its usability in the broader context of robust training in computer vision. In this paper, we build upon this work by obtaining an upper bound for attributional vulnerability as a function of spatial correlation between the input image and its explanation map. Furthermore, we also introduce a training technique that minimizes this upper bound to provide attributional robustness. In particular, we introduce a training methodology for attributional robustness that uses soft-margin triplet loss to increase the spatial correlation of input with its attribution map. The triplet loss considers input image as the anchor, gradient of the correct class logit with respect to input as the positive and gradient of the incorrect class with highest logit value with respect to input as the negative. We show empirically how this choice results in learning of robust and interpretable features that help in other downstream weakly supervised tasks.

Existing related efforts in deep learning research are largely focused on robustness to adversarial perturbations [17, 55], which are imperceptible perturbations which, when added to input, drastically change the neural network's prediction. While adversarial robustness has been explored significantly in re-

cent years, there has been limited progress made on the front of attributional robustness, which we seek to highlight in this work. Our main contributions can be summarized as:

- We tackle the problem of attribution vulnerability and provide an upper bound for it as a function of spatial correlation between the input and its attribution map [48]. We then propose *ART*, a new training method that aims to minimize this bound to learn attributionally robust model.
- Our method outperforms prior work and achieves state-of-the-art attributional robustness on Integrated Gradient [54] based attribution method.
- We show that the proposed methodology also induces immunity to adversarial perturbations and common perturbations [20] on standard vision datasets that is comparable to the state-of-the-art adversarial training technique [31].
- We show the utility of *ART* for other computer vision tasks such as weakly supervised object localization (WSOL) and segmentation. Specifically, *ART* achieves state-of-the-art performance in WSOL task on CUB-200 [61] dataset.

2 Related Work

Our work is associated with various recent development made in the field of explanation methods, robustness to input distribution shifts and weakly supervised object localization. We hence describe earlier efforts in these directions below.

Visual Explanation Methods: Various explanation methods have been proposed that focus on producing posterior explanations for the model's decisions. A popular approach to do so is to attribute the predictions to the set of input features [48, 52, 47, 54, 46, 6]. [69, 13] provide a survey of interpretation techniques. Another class of explanation methods, commonly referred to as attribution techniques, can be broadly divided into three categories - gradient/back-propagation, propagation and perturbation based methods. Gradient-based methods attribute an importance score for each pixel by using the derivative of a class score with respect to input features [48, 47, 54]. Propagation-based techniques [6, 46, 67] leverage layer-wise propagation of feature importance to calculate the attribution maps. Perturbation-based interpretation methods generate attribution maps by examining the change in prediction of the model when the input image is perturbed [65, 40, 41]. In this work, we primarily report results on the attribution method of Integrated Gradients *IG* [54] that satisfies desirable axiomatic properties and was also used in the previous work [10].

Robustness of Attribution Maps: Recently, there have been a few efforts [70, 16, 12, 10, 3] that have explored the robustness of attribution maps, which we call attributional robustness in this work. The authors of [16, 12, 70] study the robustness of a network's attribution maps and show that the attribution maps can be significantly manipulated via imperceptible input perturbations while preserving the classifier's prediction. Recently, Chen, J. et al.[10] proposed a robust attribution training methodology, which is one of the first attempts at

making an image classification model attributionally robust and is the current state of the art. The method minimizes the norm of difference in Integrated Gradients [54] of an original and perturbed image during training. In this work, we approach the problem from a different perspective of maintaining spatial alignment between an image and its saliency map.

Adversarial Perturbation and Robustness: Adversarial attacks can be broadly categorized into two types: White-box [33, 31, 8, 62] and Black-box attacks [22, 58, 2, 39]. Several proposed defense techniques have been shown to be ineffective to adaptive adversarial attacks [5, 28, 8, 7]. Adversarial training [18, 31, 50], which is a defense technique that continuously augments the data with adversarial examples while training, is largely considered the current state-of-the-art to achieve adversarial robustness. [66] characterizes the trade-off between accuracy and robustness for classification problems and propose a regularized adversarial training method. Prior works have also attempted to improve adversarial robustness using gradient regularization that minimizes the Frobenius norm of the Hessian of the classification loss with respect to input[42, 34, 30] or weights [23]. For a comprehensive review of the work done in the area of adversarial examples, please refer [63, 1].

We show in our work that in addition to providing attributional robustness, our proposed method helps in achieving performance gain on downstream task of WSOL. We hence briefly discuss earlier efforts on this task below.

Weakly Supervised Object Localization (WSOL): The problem of WSOL aims to identify the location of the object in a scene using only image-level labels, and without any location annotations. Generally, rich labeled data is scarcely available, and its collection is expensive and time-consuming. Learning from weak supervision is hence promising as it requires less rich labels and has the potential to scale. A common problem with most previous approaches is that the model only identifies the most discriminative part of the object rather than the complete object. For example, in the case of a bird, the model may rely on the beak region for classification than the entire bird's shape. In WSOL task, ADL [11], the current state-of-the-art method, uses an attention-based dropout layer while training the model that promotes the classification model to also focus on less discriminative parts of the image. For getting the bounding box from the model, ADL and similar other techniques in this domain first extract attribution maps, generally CAM-based[71], for each image and then fit a bounding box as described in [71]. We now present our methodology.

3 Attributional Robustness Training: Methodology

Given an input image $x \in [0,1]^n$ with true label $y \in \{1...k\}$, we consider a neural network model $f_\theta : \mathbb{R}^n \to \mathbb{R}^k$ with ReLU activation function that classifies x into one of k classes as $\arg\max f(x)_i$ where $i \in \{1...k\}$. Here, $f(x)_i$ is the i^{th} logit of $f(x)$. Attribution map $A(x, f(x)_i) : \mathbb{R}^n \to \mathbb{R}^n$ with respect to a given

class i assigns an importance score to each input pixel of x based on its relevance to the model for predicting the class i.

3.1 Attribution Manipulation

It was shown recently [12, 16] that for standard models f_θ, it is possible to manipulate the attribution map $A(x, f(x)_y)$ (denoted as $A(x)$ for simplicity in the rest of the paper) with visually imperceptible perturbation δ in the input by optimizing the following loss function.

$$\begin{aligned} &\arg\max_{\delta \in B_\epsilon} D[A(x+\delta, f(x+\delta)_y), A(x, f(x)_y)] \\ &\text{subject to: } \arg\max(f(x)) = \arg\max(f(x+\delta)) = y \end{aligned} \tag{1}$$

where B_ϵ is an l_p ball of radius ϵ centered at x and D is a dissimilarity function to measure the change between attribution maps. The manipulation was shown for various perturbation-based and gradient-based attribution methods.

This vulnerability in neural network-based classification models suggests that the model relies on features different from what humans perceive as important for its prediction. The goal of attributional robustness is to mitigate this vulnerability and ensure that attribution maps of two visually indistinguishable images are also nearly identical. In the next section, we propose a new training methodology for attributional robustness motivated from the observation that feature importance in image space has a high spatial correlation with the input image for robust models [57, 15].

3.2 Attributional Robustness Training (*ART*)

Given an input image $x \in \mathbb{R}^n$ with ground truth label $y \in \{1...k\}$ and a classification model f_θ, the gradient-based feature importance score is defined as $\nabla_x f(x)_i : i \in \{1...k\}$ and denoted as $g^i(x)$ in the rest of the paper. For achieving attributional robustness, we need to minimize the attribution vulnerability to attacks as defined in Equation 1. Attribution vulnerability can be formulated as the maximum possible change in $g^y(x)$ in a ϵ-neighborhood of x if A is taken as gradient attribution method [48] and D is a distance measure in some norm $||.||$ i.e.

$$\max_{\delta \in B_\epsilon} ||g^y(x+\delta) - g^y(x)|| \tag{2}$$

We show that Equation 2 is upper bounded by the maximum of the distance between $g^y(x+\delta)$ and $x+\delta$ for δ in ϵ neighbourhood of x.

$$\begin{aligned} ||g^y(x+\delta) - g^y(x)|| &= ||g^y(x+\delta) - (x+\delta) - (g^y(x) - x) + \delta|| \\ &\leq ||g^y(x+\delta) - (x+\delta)|| + ||g^y(x) - x|| + ||\delta|| \\ &\leq ||g^y(x+\delta) - (x+\delta)|| + \max_{\delta \in B_\epsilon} ||g^y(x+\delta) - (x+\delta)|| + ||\delta|| \end{aligned} \tag{3}$$

Taking max on both sides:

$$\max_{\delta \in B_\epsilon} ||g^y(x+\delta) - g^y(x))|| \leq 2 \max_{\delta \in B_\epsilon} ||g^y(x+\delta) - (x+\delta)|| + ||\epsilon|| \tag{4}$$

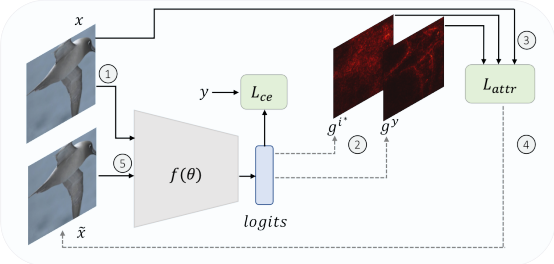


Fig. 2: Block diagram summarizing our training technique for *ART*. Dashed line represents backward gradient flow, and bold lines denotes forward pass of the neural network.

Leveraging existing understanding [44, 21] that minimizing the distance between two quantities can benefit from a negative anchor, we use a triplet loss formulation as defined in Equation 5 with image x as an anchor, $g^y(x)$ as positive sample and $g^{i^*}(x)$ as negative sample. More details about the selection of the optimization objective 5 and choice for the negative sample can be found in the supplementary section 1.1. Hence to achieve attributional robustness, we propose a training technique *ART* that encourages high spatial correlation between $g^y(x)$ and x by optimizing L_{attr} which is a triplet loss [21] with soft margin on cosine distance between $g^i(x)$ and x i.e.

$$\begin{aligned} L_{attr}(x,y) &= \log\Big(1+\exp\big(-(d(g^{i^*}(x),x)-d(g^y(x),x))\big)\Big) \\ \text{where } d(g^i(x),x) &= 1-\frac{g^i(x).x}{||g^i(x)||_2.||x||_2}\ ;\ \ i^* = \arg\max_{i\neq y} f(x)_i \end{aligned} \tag{5}$$

Hence, the classification training objective for *ART* methodology is:

$$\begin{aligned} &\underset{\theta}{\text{minimize}}\ \mathop{\mathbb{E}}_{(x,y)}\Big[L_{ce}(x+\delta,y)+\lambda\, L_{attr}(x+\delta,y)\Big] \\ &\text{where } \delta = \underset{||\delta||_\infty<\epsilon}{\arg\max}\ L_{attr}(x+\delta,y) \end{aligned} \tag{6}$$

Here L_{ce} is the standard cross-entropy loss. The optimization of L_{attr} involves computing gradient of $f(x)_i$ with respect to input x which suffers from the problem of vanishing second derivative in case of ReLU activation, i.e. $\partial^2 f_i/\partial x^2 \approx 0$. To alleviate this, following previous works [12, 10], we replace ReLU with softplus non-linearities while optimizing L_{attr} as it has a well-defined second derivative. The softplus approximates to ReLU as the value of β in $softplus_\beta(x) = \frac{log(1+e^{\beta x})}{\beta}$ increases. Note that optimization of L_{ce} follows the usual ReLU activation pathway. Thus, our training methodology consists of two steps: first, we calculate a perturbed image $\tilde{x} = x + \delta$ that maximizes L_{attr} through iterative projected gradient descent; secondly, we use $\tilde{x}$ as the training point on which L_{ce} and L_{attr} is minimized with their relative weightage controlled by the hyper-parameter λ.

Note that the square root of cosine distance for unit l_2 norm vectors as used in our formulation of L_{attr} is a valid distance metric and is related to the Euclidean

distance. Details about this can be found in the supplementary section 1.2. Through experiments, we empirically show that minimizing the upper bound in Equation 4 as our training objective increases the attributional robustness of the model by a significant margin. The block diagram for our training methodology is shown in Fig 2, and its pseudo-code is given in Algorithm 1.

3.3 Connection to Adversarial Robustness

For a given input image x, an adversarial example is a slightly perturbed image x' such that $||x - x'||$ is small in some norm but the model f_θ classifies x' incorrectly. Adversarial examples are calculated by optimizing a loss function L which is large when $f(x) \neq y$:

$$x_{adv} = \underset{x':||x'-x||_p<\epsilon}{\arg\max} \; L(\theta, x', y) \tag{7}$$

where L can be the cross-entropy loss, for example. For an axiomatic attribution function A which satisfies the completeness axiom i.e. $\sum_{j=1}^{n} A(x)_j = f(x)_y$, it can be shown that $|f(x)_y - f(x')_y| < ||A(x) - A(x')||_1$, as below:

$$\begin{aligned} |f(x)_y - f(x')_y| =& |\textstyle\sum_{j=1}^{n} A(x)_j - \sum_{j=1}^{n} A(x')_j| \\ \leq& \textstyle\sum_{j=1}^{n} |A(x)_j - A(x')_j| \\ =& ||A(x) - A(x')||_1 \end{aligned} \tag{8}$$

The above relationship connects adversarial robustness to attributional robustness as the maximum change in $f(x)_y$ is upper bounded by the maximum change in attribution map of x in its ϵ neighborhood. Also, it was shown [57] recently that for an adversarially robust model, gradient-based feature importance map $g^y(x)$ has high spatial correlation with the image x and it highlights the perceptually relevant features of the image. For classifiers with a locally affine approximation like a DNN with ReLU activations, Etmann et al.[15] establish theoretical connection between adversarial robustness, and the correlation of $g^y(x)$ with image x. [15] shows that for a given image x, its distance to the nearest distance boundary is upper-bounded by the dot product between x and $g^y(x)$. The authors of [15] showed that increasing adversarial robustness increases the correlation between $g^y(x)$ and x. Moreover, this correlation is related to the increase in attributional robustness of model as we show in Section 3.2.

3.4 Downstream Task: Weakly supervised Object localization (WSOL)

As an additional benefit of our approach, we show its improved performance on a downstream task - Weakly supervised Object localization (WSOL), in this case. The problem of WSOL deals with detecting objects where only class label information of images is available, and the ground truth bounding box location is inaccessible. Generally, the pipeline for obtaining bounding box locations in

Algorithm 1: Attributional Robustness Training (ART)

Input: Classification model f_θ, training data $X = \{(x_i, y_i)\}$, batch size b, number of epochs E, number of attack steps a, step-size for iterative perturbation α, softplus parameter β, weight of L_{attr} loss λ.
for *epoch* $\in \{1, 2, ..., E\}$ **do**
  Get mini-batch $x, y = \{(x_1, y_1)...(x_b, y_b)\}$
  $\tilde{x} = x + Uniform[-\epsilon, +\epsilon]$
  for *i=1,2, ... , a* **do**
    $\tilde{x} = \tilde{x} + \alpha * sign(\nabla_x L_{attr}(\tilde{x}, y))$
    $\tilde{x} = Proj_{\ell_\infty}(\tilde{x})$
  end
  $i^* = \arg\max_{i \neq y} f(x)_i$
  Calculate $g^y(\tilde{x}) = \nabla_x f(\tilde{x})_y$
  Calculate $g^{i^*}(\tilde{x}) = \nabla_x f(\tilde{x})_{i^*}$; `// We calculate` $g^y(\tilde{x})$ `and` $g^{i^*}(\tilde{x})$ `using` $softplus_\beta$ `activation as described in Section 3.2`
  $loss = L_{ce}(\tilde{x}, y) + \lambda \cdot L_{attr}(\tilde{x}, y)$
  Update θ using $loss$
end
return f_θ.

WSOL relies on attribution maps. Also, the task of object detection is widely used to validate the quality of attribution maps empirically. Since our proposed training methodology ART promotes attribution map to be invariant to small perturbations in input, it leads to better attribution maps identifying the complete object instead of focusing on only the most discriminative part of the object. We validate this empirically by using attribution maps obtained from our model for bounding-box detection on the CUB dataset and obtaining new state-of-the-art localization results.

4 Experiments and Results

In this section, we first describe the implementation details of ART and evaluation setting for measuring the attributional and adversarial robustness. We then show the performance of ART on the downstream WSOL task.

4.1 Attributional and Adversarial Robustness

Baselines: We compare our training methodology with the following approaches:

- *Natural*: Standard training with cross entropy classification loss.
- *PGD-n*: Adversarially trained model with n-step PGD attack as in [31], which is typically used by work in this area [10].
- *IG Norm* and *IG-SUM Norm* [10]: Current state-of-the-art robust attribution training technique.

Table 1: Attributional and adversarial robustness of different approaches on various datasets. Hyper-parameters for attributional attack are same as [10]. Similarity measures used are IN: *Top-k intersection*, K:*kendall's tau rank order correlation*. The values denote similarity between attribution maps of original and perturbed examples [16] based on *Intergrated Gradient* method.

Dataset	**Approach**	**Attributional Robustness**		**Accuracy**	
		IN	K	Natural	PGD-40 Attack
CIFAR-10	Natural	40.25	49.17	95.26	0.
	PGD-10 [31]	69.00	72.27	87.32	44.07
	ART	**92.90**	**91.76**	89.84	37.58
SVHN	Natural	60.43	56.50	95.66	0.
	PGD-7 [31]	39.67	55.56	92.84	50.12
	ART	**61.37**	**72.60**	95.47	43.56
GTSRB	Natural	68.74	76.48	99.43	19.9
	IG Norm [10]	74.81	75.55	97.02	75.24
	IG-SUM Norm [10]	74.04	76.84	95.68	77.12
	PGD-7 [31]	86.13	88.42	98.36	87.49
	ART	**91.96**	**89.34**	98.47	84.66
Flower	Natural	38.22	56.43	93.91	0.
	IG Norm [10]	64.68	75.91	85.29	24.26
	IG-SUM Norm [10]	66.33	79.74	82.35	47.06
	PGD-7 [31]	**80.84**	84.14	92.64	69.85
	ART	79.84	**84.87**	93.21	33.08

Datasets and Implementation Details: To study the efficacy of our methodology, we benchmark on the following standard vision datasets: CIFAR-10 [27], SVHN [35], GTSRB [53] and Flower [36]. For CIFAR-10, GTSRB and Flower datasets, we use Wideresnet-28-10 [64] model architecture for *Natural*, *PGD-10* and *ART*. For SVHN, we use WideResNet-40-2 [64] architecture. We use the perturbation $\epsilon = 8/255$ in ℓ_∞-norm for *ART* and *PGD-n* as in [31, 10]. We use $\lambda = 0.5$, $a = 3$ and $\beta = 50$ for all experiments in the paper. For training, we use SGD optimizer with step-wise learning rate schedule. More details about training hyper-parameters can be found in the supplementary section 1.3.

Evaluation: For evaluating attributional robustness, we follow [10] and present our results with Integrated Gradient (IG)-based attribution maps. We show attributional robustness of *ART* on other attribution methods in Section 5. IG satisfies several theoretical properties desirable for an attribution method, e.g. sensitivity and completeness axioms and is defined as:

$$IG(x, f(x)_i) = (x - \overline{x}) \odot \int_{t=0}^{1} \nabla_x f(\overline{x} + t(x - \overline{x}))_i dt \tag{9}$$

where $\overline{x}$ is a suitable baseline at which the function prediction is neutral. For computing perturbed image $\tilde{x}$ on which $IG(\tilde{x})$ changes drastically from $IG(x)$, we perform Iterative Feature Importance Attack (IFIA) proposed by Ghorbani et al.[16] with ℓ_∞ bound of $\epsilon = 8/255$ as used by previous work [10].

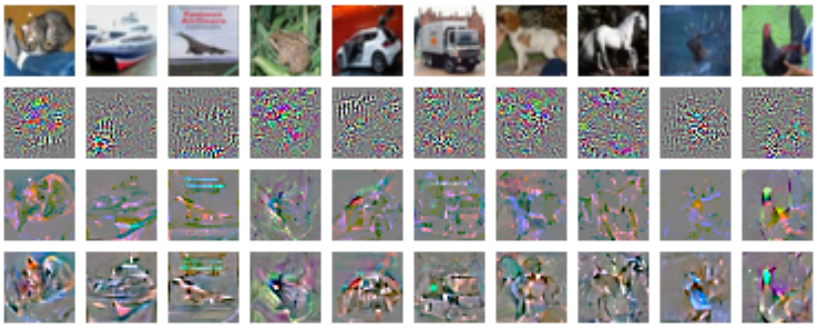


Fig. 3: Examples of gradient attribution map [48] for different models on CIFAR-10. Top to bottom: Image; attribution maps for *Natural*, *PGD-10* and *ART* models

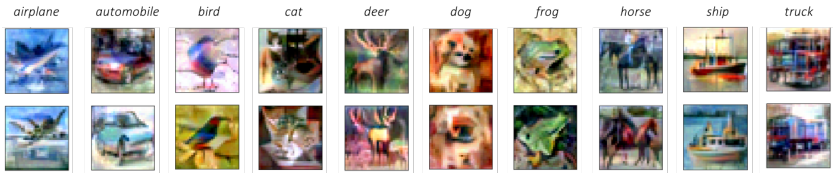


Fig. 4: Random samples (of resolution 32×32) generated using a CIFAR-10 robustly trained *ART* classifier

For assessing similarity between $A(x)$ and perturbed image $A(\tilde{x})$, we use *Top-k intersection* (IN) and *Kendall's tau coefficient* (K) similar to [10]. *Kendall's tau coefficient* is a measure of similarity of ordering when ranked by values, and therefore is a suitable metric for comparing attribution maps. *Top-k* intersection measures the percentage of common indices in top-k values of attribution map of x and $\tilde{x}$. We report average of IN and K metric over random 1000 samples of test-set. More details about the attack methodology and evaluation parameters can be found in supplementary section 1.3. For evaluating adversarial robustness, we perform 40 step PGD attack [31] using cross-entropy loss with ℓ_∞ bound of $\epsilon = 8/255$ and report the model accuracy on adversarial examples. Table 1 compares attributional and adversarial robustness across different datasets and training approaches. *ART* achieves state-of-the-art attributional robustness on attribution attacks [16] when compared with baselines. We also observe that *ART* consistently achieves higher test accuracy than [31] and has adversarial robustness significantly greater than that of the *Natural* model.

Qualitative study of input-gradients for ART: Motivated by [57] which claims that adversarially trained models exhibits human-aligned gradients (agree with human saliency), we studied the same with (*ART*), and the results are shown in Fig 3. Qualitative study of input-gradients shows a high degree of spatial alignment between the object and the gradient. We also show image generation from random seeds in Fig 4 using robust ART model as done in [43]. The image generation process involves maximization of the class score of the

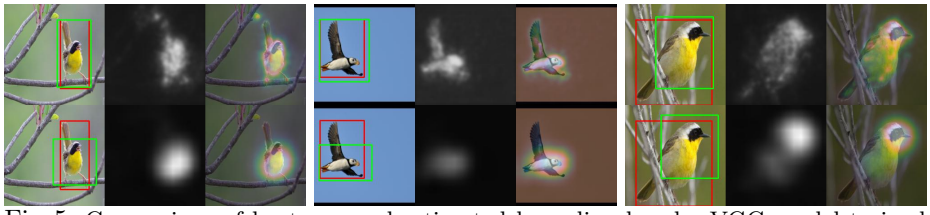


Fig. 5: Comparison of heatmap and estimated bounding box by VGG model trained via ART (top row) and ADL (bottom row) on CUB dataset; The red bounding box is ground truth and green bounding box corresponds to the estimated box

Table 2: Weakly Supervised Localization on CUB dataset. Bold text refers to the best GT-Known Loc and Top-1 Loc for each model. $*$ denotes directly reported from the paper. # denotes our implementation from the official code released by ADL [11][2]

Model	Method	**Saliency Method**				Top-1 Acc
		Grad		CAM		
		GT-Known Loc	Top-1 Loc	GT-Known Loc	Top-1 Loc	
ResNet50-SE	ADL [11]	-	-	-	62.29*	80.34*
ResNet50	ADL#	52.93	43.78	56.85	47.53	80.0
	Natural	50.2	42.0	60.37	50.0	81.12
	PGD-7[31]	66.73	47.48	55.24	39.45	70.3
	ART	**82.65**	**65.22**	58.87	46.02	77.51
VGG-GAP	ADL#	63.18	43.59	69.36	50.88	70.31
	Natural	72.54	53.81	48.75	35.03	72.94
	ART	**76.50**	**57.74**	52.88	40.75	74.51

desired class starting from a random seed which is sampled from some class-conditional seed distribution as defined in [43].

4.2 Weakly Supervised Image Localization

This task relies on the attribution map obtained from the classification model to estimate a bounding box for objects. We compare our approach with ADL [11][3] on the CUB dataset, which has ground truth bounding box of 5794 bird images. We adopt similar processing steps as ADL for predicting bounding boxes except that we use gradient attribution map $\nabla_x f(x)_y$ instead of CAM [71]. As a post-processing step, we convert the attribution map to grayscale, normalize it and then apply a mean filtering of 3×3 kernel over it. Then a bounding box is fit over this heatmap to localize the object.

We perform experiments on Resnet-50 [19] and VGG [49] architectures. We use ℓ_∞ bound of $\epsilon = 2/255$ for *ART* and *PGD-7* training on CUB dataset. For evaluation, we used similar metrics as in [11] i.e. *GT-Known Loc*: Intersection over Union (IoU) of estimated box and ground truth bounding box is atleast 0.5 and ground truth is known; *Top-1 Loc*: prediction is correct and IoU of bounding box is atleast 0.5; *Top-1 Acc*: top-1 classification accuracy. More details about

[3] `https://github.com/junsukchoe/ADL/tree/master/Pytorch`

Table 3: Top-1 accuracy of different models on perturbed variants of test-set (GN:Gaussian noise; SN: Shot noise; IN: Impulse noise; DB: Defocus blur; Gl-B: Glass blur; MB: Motion blur; ZB: Zoom blur; S: Snow; F: Fog; B: Brightness; C: Contrast; E: Elastic transform; P: Pixelation noise; J: JPEG compression; Sp-N: Speckle Noise)

Models	**GN**	**SN**	**IN**	**DB**	**Gl-B**	**MB**	**ZB**	**S**	**F**	**B**	**C**	**E**	**P**	**J**	**Sp-N**
Natural	49.16	61.42	59.22	83.55	53.84	79.16	79.18	84.53	**91.6**	**94.37**	**87.63**	84.44	74.12	79.76	65.04
PGD-10	83.32	84.33	73.73	83.09	81.27	79.60	82.07	82.68	68.81	85.97	57.86	81.68	85.56	85.56	83.64
ART	**85.44**	**86.41**	**77.07**	**86.07**	**81.70**	**83.14**	**85.54**	**84.99**	71.04	89.42	56.69	**84.72**	**87.64**	**87.89**	**86.02**

Table 4: Attributional Robustness on CIFAR-10 for other attribution methods

Model	**Gradient[48]**		**GradSHAP [29]**	
	IN	K	IN	K
Natural	13.72	9.5	4.5	16.52
PGD-10 [31]	54.8	54.06	45.05	59.80
ART	**76.07**	**70.31**	**48.31**	**62.35**

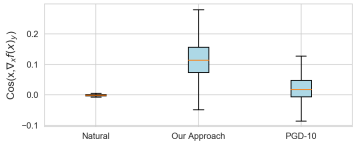


Fig. 6: Cosine(x,$\nabla_x f(x)_y$) for different models over test-set of CIFAR-10

dataset and hyper-parameters can be found in the supplementary section 2.1. Our approach results in higher *GT-Known Loc* and *Top-1 Loc* for both Resnet-50 and VGG-GAP [11] model as shown in Table 2. We also show qualitative comparison of the bounding box estimated by our approach with [11] in Fig 5.

5 Discussion and Ablation Studies

To understand the scope and impact of the proposed training approach *ART*, we perform various experiments and report these findings in this section. These studies were carried out on the CIFAR-10 dataset.

Robustness to targeted attribution attacks: In targeted attribution attacks, the aim is to calculate perturbations that minimize dissimilarity between the attribution map of a given image and a target image's attribution map. We evaluate the robustness of *ART* model using targeted attribution attack as proposed in [12] using the *IG* attribution method on a batch of 1000 test examples. To obtain the target attribution maps, we randomly shuffle the examples and then evaluate *ART* and *PGD-10* trained model on these examples. The *kendall's tau coefficient* and *top-k intersection* similarity measure between original and perturbed images on *ART* was 64.76 and 70.64 as compared to 36.29 and 31.81 on the *PGD-10* adversarially trained model.

Attributional robustness for other attribution methods: We evaluate *ART* against attribution attack [16] using gradient[48] and gradSHAP[29] attribution methods in Table 4. We observe that *ART* achieves higher attributional robustness than *Natural* and *PGD-10* models on *Top-k intersection* (IN) and *Kendall's tau coefficient* (K) measure. We also compare the cosine similarity between x and $g^y(x)$ for all models trained on CIFAR-10 dataset and show its variance plot in Fig. 6. We can see that *ART* trained model achieves higher

cosine similarity than *Natural* and *PGD-10*. This empirically validates that our optimization is effective in increasing the spatial correlation between x and $g^y(x)$.

Robustness against gradient-free and stronger attacks: To show the absence of gradient masking and obfuscation [5, 7], we evaluate our model on a gradient-free adversarial optimization algorithm [58] and a stronger PGD attack with a larger number of steps. We observe similar adversarial robustness when we increase the number of steps in PGD-attack. For 100 step and 500 step PGD attacks, *ART* achieves 37.42 % and 37.18 % accuracy respectively. On the gradient-free SPSA [58] attack, *ART* obtains 44.7 adversarial accuracy that was evaluated over 1000 random test samples.

Robustness to common perturbations [20] and spatial adversarial perturbations [14]: We compare *ART* with *PGD-10* adversarially trained model on the common perturbations dataset [20] for CIFAR-10. The dataset consists of perturbed images of 15 common-place visual perturbations at five levels of severity, resulting in 75 distinct corruptions. We report the mean accuracy over severity levels for all 15 types of perturbations and observe that *ART* performs better than other models on a majority of these perturbations, as shown in Table 3. On PGD-40 ℓ_2 norm attack with $\epsilon = 1.0$ and spatial attack [14] we observe robustness of 39.65%, 11.13% for *ART* and 29.68%, 6.76% for *PGD-10* trained model, highlighting the improved robustness provided by our method. More results of varying ϵ in adversarial attacks and combining *PGD* adversarial training [31] with *ART* can be found in the supplementary section 3.

Image Segmentation: Data collection for image segmentation task is time-consuming and costly. Hence, recent efforts [26, 59, 60, 25, 38, 68, 37] have focused on weakly supervised segmentation models, where image labels are leveraged instead of segmentation masks. Since models trained via our approach perform well on WSOL, we further evaluate it on weakly supervised image segmentation task for Flowers dataset [36] where we have access to segmentation masks of 849 images. Samples of weakly-supervised segmentation mask obtained from attribution maps on various models are shown in Fig. 7. We observe that attribution maps of *ART* can serve as a better prior for segmentation masks as compared to other baselines. We evaluate our results using *Top-1 Seg* metric which considers an answer as correct when the model prediction is correct and the IoU betweeen ground-truth mask and estimated mask is atleast 0.5. We compare *ART* against *Natural* and *PGD-7* trained models using gradient[48] and IG [54] attribution maps. Attribution maps are converted into gray-scale heatmaps and a smoothing filter is applied as a post-processing step. We obtain a *Top-1 Seg* performance of 0.337, 0.422, and 0.604 via IG attribution maps and 0.244, 0.246, 0.317 via gradient maps for *Natural*, *PGD-7* and *ART* respectively.

Effect of β, λ and a on performance: We perform experiments to study the role of β, λ and a as used in Algorithm 1 on the model performance by varying one parameter and fixing the others on their best-performing values, i.e. 50, 0.5 and 3 respectively. Fig. 8a shows the plots of attributional robustness. Fig. 8b shows the plots of test accuracy and adversarial accuracy on ℓ_∞ PGD-40 perturbations with $\epsilon = 8/255$. We observe that adversarial and attributional robustness

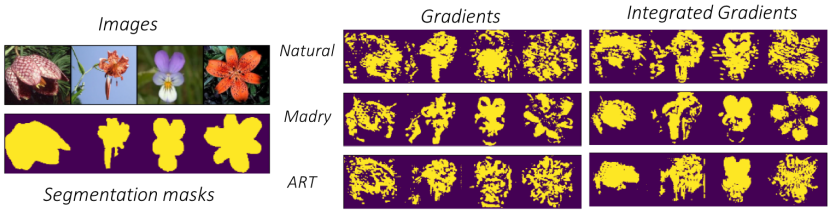
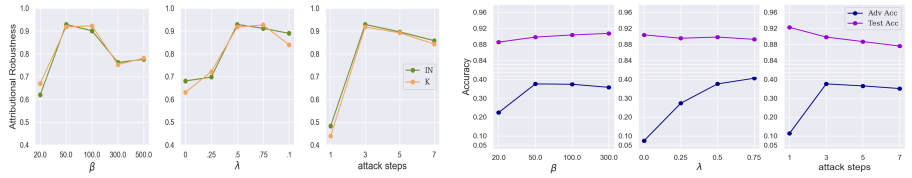


Fig. 7: Example images of weakly supervised segmentation masks obtained from different models via different attribution methods



(a) (b)

Fig. 8: (a): Top-k Intersection (IN) and Kendall correlation (K) measure of attributional robustness; (b): Test accuracy and adversarial accuracy (PGD-40 perturbations) on varying β, λ and attack steps in our training methodology on CIFAR-10

initially increases with increasing β, but the trend reverses for higher values of β. On varying λ, we find that the attributional and adversarial robustness of the model increases with increasing λ and saturates after 0.75. For attack steps parameter a, we find that the performance in terms of test accuracy, adversarial accuracy and attributional robustness saturates after 3 attack steps as shown in the right-most plot of Fig. 8a and 8b.

6 Conclusion

We propose a new method for the problem space of attributional robustness, using the observation that increasing the alignment between the object in an input and the attribution map generated from the network's prediction leads to improvement in attributional robustness. We empirically showed this for both un-targeted and targeted attribution attacks over several benchmark datasets. We showed that the attributional robustness also brings out other improvements in the network, such as reduced vulnerability to adversarial attacks and common perturbations. For other vision tasks such as weakly supervised object localization, our attributionally robust model achieves a new state-of-the-art accuracy even without being explicitly trained to achieve that objective. We hope that our work can open a broader discussion around notions of robustness and the application of robust features on other downstream tasks.

Acknowledgements. This work was partly supported by the Ministry of Human Resource Development and Department of Science and Technology, Govt of India through the UAY program.

References

1. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access (2018)
2. Alexey Kurakin, Ian J. Goodfellow, S.B.: Adversarial examples in the physical world. ICLR Workshop (2017)
3. Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods. ICML 2018 Workshop (2018)
4. Ardila, D., Kiraly, A.P., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D.P., Shetty, S.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine **25**, 954—-961 (2019)
5. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. ICML (2018)
6. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müler, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7): e0130140 (2015)
7. Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., Kurakin, A.: On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705 (2019)
8. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP) (2017)
9. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. arXiv preprint arXiv:1710.11063 (2017)
10. Chen, J., Wu, X., Rastogi, V., Liang, Y., Jha, S.: Robust attribution regularization. arXiv preprint arXiv:1905.09957 (2019)
11. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2219–2228 (2019)
12. Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame. In: Advances in Neural Information Processing Systems. pp. 13567–13578 (2019)
13. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. arXiv preprint arXiv:1808.00033 (2018)
14. Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., Madry, A.: Exploring the landscape of spatial robustness. In: International Conference on Machine Learning. pp. 1802–1811 (2019)
15. Etmann, C., Lunz, S., Maass, P., Schönlieb, C.B.: On the connection between adversarial robustness and saliency map interpretability. arXiv preprint arXiv:1905.04172 (2019)
16. Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3681–3688 (2019)
17. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015)
18. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. ICLR (2015)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)

20. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (2019)
21. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
22. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box adversarial attacks with limited queries and information. ICML (2018)
23. Jakubovitz, D., Giryes, R.: Improving dnn robustness to adversarial attacks using jacobian regularization. ECCV (2018)
24. Jia, X., Shen, L.: Skin lesion classification using class activation map. arXiv preprint arXiv:1703.01053 (2017)
25. Jiang, Q., Tawose, O.T., Pei, S., Chen, X., Jiang, L., Wang, J., Zhao, D.: Weakly-supervised image semantic segmentation based on superpixel region merging. Big Data and Cognitive Computing **3**(2), 31 (2019)
26. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. CoRR **abs/1603.06098** (2016), `http://arxiv.org/abs/1603.06098`
27. Krizhevsky, A., Nair, V., Hinton, G.: Cifar-10. URL `http://www.cs.toronto.edu/kriz/cifar.html` (2010)
28. Logan Engstrom, Andrew Ilyas, A.A.: Evaluating and understanding the robustness of adversarial logit pairing. NeurIPS SECML (2018)
29. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) NeurIPS (2017), `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`
30. Lyu, C., Huang, K., Liang, H.N.: A unified gradient regularization family for adversarial examples. ICDM (2015)
31. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
32. Mitani, A., Huang, A., Venugopalan, S., Corrado, G.S., Peng, L., Webster, D.R., Hammel, N., Liu, Y., Varadarajan, A.V.: Detection of anaemia from retinal fundus images via deep learning. Nature Biomedical Eng. **4**, 18–27 (2020)
33. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. arXiv preprint arXiv:1511.04599v3 (2016)
34. Moosavi-Dezfooli, S.M., Fawzi, A., Uesato, J., Frossard, P.: Robustness via curvature regularization, and vice versa. CVPR (2019)
35. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
36. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 1447–1454 (2006)
37. Nilsback, M.E., Zisserman, A.: Delving deeper into the whorl of flower segmentation. Image Vision Comput. **28**(6), 1049–1062 (Jun 2010). https://doi.org/10.1016/j.imavis.2009.10.001, `http://dx.doi.org/10.1016/j.imavis.2009.10.001`
38. Oh, S.J., Benenson, R., Khoreva, A., Akata, Z., Fritz, M., Schiele, B.: Exploiting saliency for object segmentation from image level labels. CoRR **abs/1701.08261** (2017), `http://arxiv.org/abs/1701.08261`
39. Papernot, N., McDaniel, P., Goodfellow, I.J., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. ACM (2017)

40. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: BMVC (2018)
41. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: ACM SIGKDD (2016)
42. Ross, A.S., Doshi-Velez, F.: Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. AAAI (2018)
43. Santurkar, S., Ilyas, A., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Image synthesis with a single (robust) classifier. In: NeurIPS (2019)
44. Schroff, Florian an Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. CVPR (2015)
45. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization (2016)
46. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. pp. 3145–3153 (2017)
47. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713 (2016)
48. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
49. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
50. Sinha, A., Singh, M., Kumari, N., Krishnamurthy, B., Machiraju, H., Balasubramanian, V.: Harnessing the vulnerability of latent layers in adversarially trained models. arXiv preprint arXiv:1905.05186 (2019)
51. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. Workshop on Visualization for Deep Learning, ICML (2017)
52. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. ICLR workshop (2015)
53. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In: IEEE International Joint Conference on Neural Networks. pp. 1453–1460 (2011)
54. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. ICML (2017)
55. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. ICLR (2014)
56. Taly, A., Joseph, A., Sood, A., Webster, D., Coz, D.D., Wu, D., Rahimy, E., Corrado, G., Smith, J., Krause, J., Blumer, K., Peng, L., Shumski, M., Hammel, N., Sayres, R.A., Barb, S., Rastegar, Z.: Using a deep learning algorithm and integrated gradient explanation to assist grading for diabetic retinopathy. Ophthalmology (2019)
57. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. In: ICLR (2019)
58. Uesato, J., O'Donoghue, B., Kohli, P., van den Oord, A.: Adversarial risk and the dangers of evaluating against weak attacks. ICML (2018)
59. Vasconcelos, M., Vasconcelos, N., Carneiro, G.: Weakly supervised top-down image segmentation. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 1, pp. 1001–1006 (June 2006). https://doi.org/10.1109/CVPR.2006.333

60. Vezhnevets, A., Buhmann, J.M.: Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3249–3256 (June 2010). https://doi.org/10.1109/CVPR.2010.5540060
61. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
62. Xu, K., Liu, S., Zhao, P., Chen, P.Y., Zhang, H., Fan, Q., Erdogmus, D., Wang, Y., Lin, X.: Structured adversarial attack: Towards general implementation and better interpretability. ICLR (2019)
63. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems (2019)
64. Zagoruyko, S., Komodakis, N.: Wide residual networks. CoRR **abs/1605.07146** (2016), `http://arxiv.org/abs/1605.07146`
65. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)
66. Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. arXiv preprint arXiv:1901.08573 (2019)
67. Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. ECCV (2016)
68. Zhang, L., Song, M., Liu, Z., Liu, X., Bu, J., Chen, C.: Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1908–1915 (June 2013). https://doi.org/10.1109/CVPR.2013.249
69. Zhang, Q., Zhu, S.C.: Visual interpretability for deep learning: a survey. arXiv preprint arXiv:1802.00614 (2018)
70. Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., Wang, T.: Interpretable deep learning under fire. arXiv preprint arXiv:1812.00891 (2018)
71. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)