Explaining Classifiers using Adversarial Perturbations on the Perceptual Ball

Andrew Elliott*
Uni. of Glasgow/Turing Institute
andrew.elliott@glasgow.ac.uk

Stephen Law*
UCL/Turing Institute
stephen.law@ucl.ac.uk

Chris Russell†
Amazon Tübingen
cmruss@amazon.com

Abstract

We present a simple regularization of adversarial perturbations based upon the perceptual loss. While the resulting perturbations remain imperceptible to the human eye, they differ from existing adversarial perturbations in that they are semi-sparse alterations that highlight objects and regions of interest while leaving the background unaltered. As a semantically meaningful adverse perturbations, it forms a bridge between counterfactual explanations and adversarial perturbations in the space of images.

We evaluate our approach on several standard explainability benchmarks, namely, weak localization, insertion-deletion, and the pointing game demonstrating that perceptually regularized counterfactuals are an effective explanation for image-based classifiers.

1. Introduction

We address the gap between counterfactual explanations [53] and adversarial perturbations [48], and show why minimal changes in image data that results in a change in classifier response does not result in semantically meaningful alteration. One might hope that the smallest edit to alter classifier response of an image labeled as bird should alter the bird pixels, but in practice adversarial perturbations make non-local changes that break the classifier. We show how penalizing changes in the mid-level classifier response with a perceptual loss stops this breakage and instead results in semantically meaningful changes that highlight the extent of objects in images (see Figs. 1,2).

Outside of computer vision [53], counterfactual explanations are a popular method in explainable AI. They find the smallest change needed to alter the decision of a classifier, and on tabular data, can give explanations such as:

"The loan was denied as your income was £30,000. If it had been £45,000, you would have been offered a loan."

*Equal contribution
†Much of this work was done prior to joining Amazon.

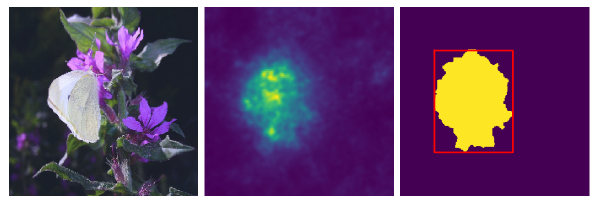

Figure 1. Object localization. From left to right: Original image; Magnitude of the perceptual perturbations; Dominant connected component and the bounding box from automatic object detection. Despite the flowers our method highlights the butterfly as salient.

The close relationship between adversarial perturbations and counterfactual explanations follows from the definitions in philosophy and folk psychology of a counterfactual explanation as answering the question "What is a minimal change that would result in a different outcome?" When applied to classifiers, these counterfactual explanations are simply adversarial perturbations, by a different name. As such, it is interesting to ask, why adversarial perturbations don't work as explanations: Why are they imperceptible? and: Why don't they localize on objects?

Two compelling arguments for the existence of imperceptible adversarial perturbations in images have been offered. The first due to [16] remarks that they are simply an artifact of a high-dimensional space, and thus, it is entirely expected that a small perturbation of every pixel can add up to a large change in the classifier response, and that in fact the same behavior is found in linear classifiers.

A second argument attempts to understand why sparse (potentially single pixel) attacks exist and attributes the effectiveness of adversarial perturbations to exploding gradients. 'Exploding gradients' refers to the phenomenon where changes in functional response grow exponentially with the depth of the network, relative to a change of input of fixed magnitude. These exploding gradients are an issue known to afflict the learning of Recurrent Neural Networks [30], and the deep networks common to computer vision. This phenomenon occurs because, by construction, neural networks form a product of (convolutional) matrix operations interlaced with nonlinearities; and for directions/locations

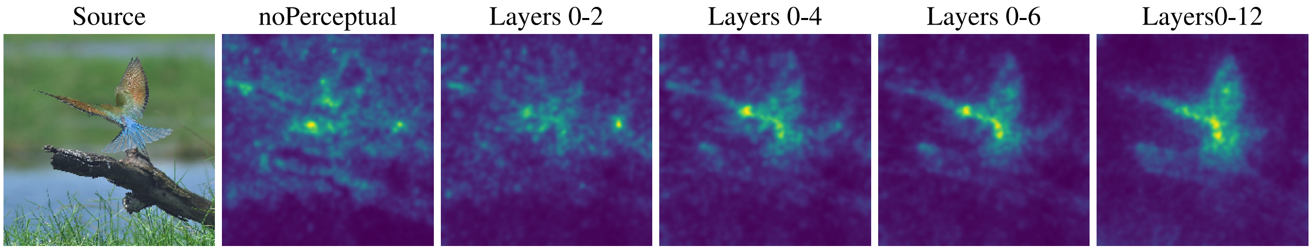

Figure 2. A representative image showing the effects of regularization over different VGG19bn layers ($\sigma = 2.0$). As we extend the regularization to cover higher layers we find the perturbation becomes more compact and better localized upon the object.

in which these nonlinearities act approximately linearly, the eigenvalues of the Jacobian can grow exponentially with depth[1]. While this phenomenon is well-studied in the context of training networks with remedies such as normalization [19] and gradient clipping [30], the same phenomenon occurs when generating adversarial perturbations. As such, a carefully chosen small perturbation to have an extremely large effect on the response of a deep or recurrent classifier.

To explore how these arguments fit together, and which explanation accounts for the familiar behavior of adversarial perturbations, we propose a simple novel regularization that bounds the exponential growth of the classifier response by regularizing the perceptual distance [20] between the image and its adversarial perturbation.

One common criticism of adversarial perturbation is that the generated images lie outside the manifold of natural images, and if we could sample from the manifold, our adversarial perturbations would be both larger and more representative of the real world. Restricting adversarial perturbations to this manifold should limit the impact of exploding gradients – if samples are drawn from this space then a well-trained classifier should implicitly reflect the smoothness of the true labels of the underlying data distribution.

While it is believed that the manifold of natural images is low dimensional [22], characterizing this manifold outside of handwritten digits has proven extremely challenging[2]. Our approach provides a complementary lightweight alternative. Rather than attempting to characterize the manifold, we penalize search directions that exploit exploding gradients as these encourage movement off the data manifold when searching for minimal adversarial perturbations.

We propose a novel regularization for adversarial perturbations based around the perceptual loss. Our new perturbations tend to highlight objects and regions of interest within the image (see Fig. 1)[3]. We evaluate on several standard explainability challenges for image classifiers and further validate using the sanity checks of [1].

[1] See [30] for a formal derivation.
[2] See discussion in the experimental section of [46].
[3] Our implementation can be found at www.github.com/alan-turing-institute/perceptualBall.

2. Prior work

Numerous approaches to adversarial perturbations have been proposed previously. These can loosely be divided into white-box [3, 6, 25, 26] approaches that assume access to the underlying nature of the model and black-box methods which do not [24, 29]. The search for an adversarial perturbation is often formulated as trying to find the closest point to a particular image, under the ℓ_∞, ℓ_1 or ℓ_2 norm that takes a different class label. Numerous defenses have been proposed [27, 36] but they can often be circumvented [49].

Other works that add additional constraints to the perturbation to try to make the generated images more plausible. Such works restrict the space of perturbations considered by trying to find an adversarial perturbation that confounds many classifiers at once [8], or is robust to image warps [2]. Other approaches considered only a single image and single classifier, but restricted adversarial perturbations to lie on the manifold of plausible images [15, 39, 43, 46]. The principal limitation of these approaches is that they require a plausible generator of natural images, something that is achievable with small simple datasets such as MNIST but currently out of reach for even the 224 by 224 thumbnails used by typical ImageNet [37] classifiers.

Adversarial Perturbations and Counterfactuals There are substantial works [52, 53] relating adversarial perturbations and counterfactual explanations. This relationship follows from the definitions in philosophy and folk psychology of a counterfactual explanation as answering the question "What could have been different in order for outcome A to have occurred instead of B?". With full causal models of images being outside our grasp, such questions are commonly answered using Lewis's Closest Possible World semantics [23], rather than Pearl's Structured Causal Models [31]. Under Lewis's framework, an explanation for why an image is classified as 'dog' rather than 'cat' can be found by searching for the most similar possible world (i.e. image) which is assigned the label 'cat' by the classifier.

Conceptually, this is no different to searching for an adversarial perturbation sampled from the space of possible images. Several approaches have been proposed that ei-

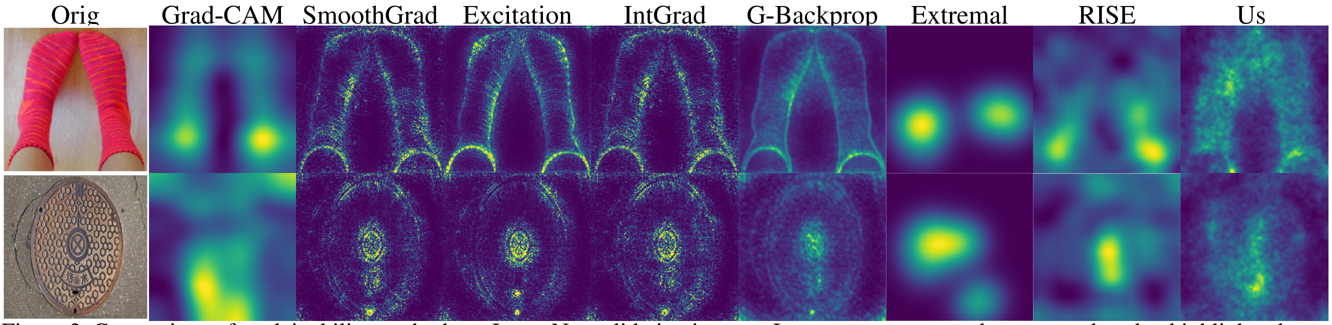

Figure 3. Comparison of explainability methods on ImageNet validation images. In contrast to many other approaches that highlight edges, ours highlights the textured interior of the object. This is consistent with the known textual bias of ImageNet trained CNNs [14].

ther bypass the requirement that the counterfactual is an image, and return text descriptions [18], naïvely ignoring the requirement that the world is plausible [53], used prototypes [51], or auto-encoders [7], or Gaussian kernels [13] to characterize the manifold of plausible images, or require large edits that replace regions of the image, either with the output of GANs [4] or with patches from other images [17].

Recent works [21, 50] suggested that robust classifiers naturally encourage perceptually aligned gradients. Our insight can be seen as complementary, rather than requiring that the classifier is robust, we prohibit perturbations from exploiting the fragility arising from exploding gradients.

Adversarial Perturbations and Gradient Methods The majority of methods in the explainability of computer vision tend to be gradient or importance-based methods that assign an importance weight to every pixel in the image; every superpixel; or to mid-level neurons. These gradient methods and adversarial perturbations are strongly related. In fact, with most modern networks being piecewise linear, if the found adversarial perturbation and the original image lie on the same linear piece, the difference between the original image and closest adversarial perturbations under the ℓ_2 norm is equivalent to the direction of steepest descent, up to scaling. As such, ℓ_2 adversarial perturbations can be thought of as a slightly robustified method of estimating the gradient, that takes into account some local non-linearities.

Of the pure gradient-based approaches, [40] calculated the output gradient with respect to the input image to create a saliency map giving fine-grained, but potentially less interpretable results. Other gradient approaches include SmoothGrad [42] which stabilizes the saliency maps by averaging over multiple noisy copies, and Integrated Gradients [47] which accumulates gradients seen when perturbing an empty image to the input image.

CAM based approaches [38, 57] sum the activation maps in the final convolutional layer of the network. These small activation maps are up-sampled to obtain a heatmap that highlights particularly salient regions. Grad-CAM is a generalized variant which finds similar regions of interest to the perturbation based approaches [38]. Recently, [34] introduced a framework that tries to unify these various gradient approach by proposing NormGrad, which aggregates the spatial gradient contributions of individual layers.

Perturbation methods estimate the local sensitivity over a larger range than gradient methods. For example, [55] applied constant occlusion masks to different input patches repeatedly to find sensitive regions. LIME [35] constructed a linear model using the responses obtained from perturbing super-pixels. The recent work on Extremal Perturbation [10] estimates an optimal mask of the image to occlude which gives a maximal effect on the network's output.

Various experiments have been proposed to test explanations including the pointing game [33, 38, 56], the weakly supervised object localization task [5, 12] and the insertion and deletion game [33, 54]. In particular, [1] developed experiments to test the suitability of saliency methods. A number of existing saliency techniques have been evaluated using these experiments, including: NormGrad [34], Extremal Perturbation [10], Gradient [40], RISE [33], Grad-CAM [38], SmoothGrad [42], GuidedBackprop [45], Integrated Gradients [47], Deconvolution [55], and Excitation Backpropagation [56]. We evaluate our approach on all these tests and compare against standard methods.

3. Methodology

We consider a classifier $C(\cdot)$ that takes an image x as input, and returns a k dimensional confidence vector.

For classifiers that assign a single class to each image, we assume the classifier $C(\cdot)$ assigns the label $i = \arg\max_j C_j(x)$ to the image x. Given image x classified as label i we consider the scalar multi-class margin:

$$M_i(x') = C_i(x') - \max_{j \neq i} C_j(x') \tag{1}$$

and note that $M_i(x') \leq 0$ if and only if $C(\cdot)$ does not assign label i to image x'.

For classifiers $C(\cdot)$ that assign multiple classes to a single image (e.g. pointing game (Sec. 4.4)), we assume that the classifier $C(\cdot)$ assigns the labels $I = \{\forall j : C_j(x) > 0\}$

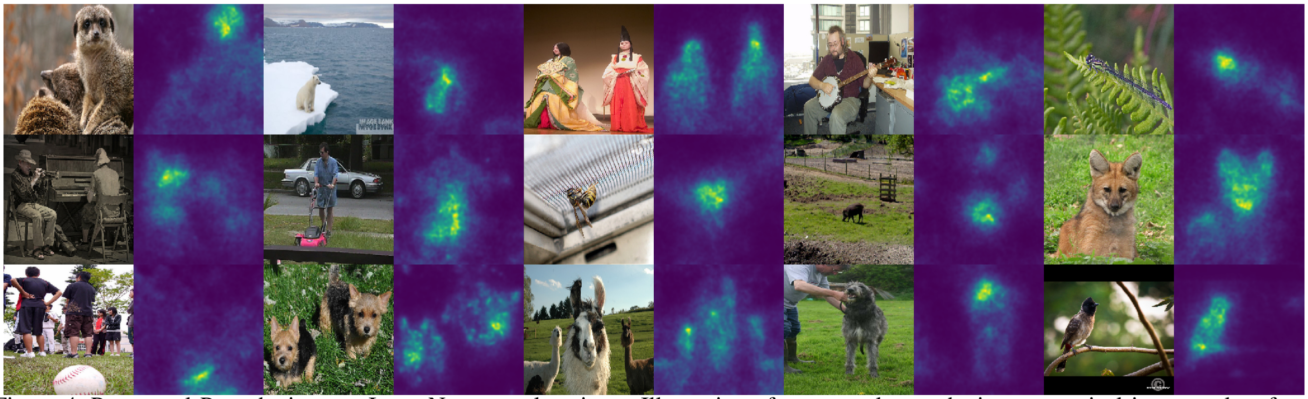

Figure 4. Perceptual Perturbations on ImageNet as explanations. Illustration of perceptual perturbations on typical images taken from ImageNet [37]. See discussion in Section 4, and for further examples see Fig. S1 in supplementary materials section A.

to the image x. For each $i \in I$, we are interested in the per label classifier response, and instead define the margin as:

$$M_i(x') = C_i(x') \tag{2}$$

Again, $M_i(x') \leq 0$ if and only if $C(\cdot)$ does not assign label i to image x. In both cases, an adversarial perturbation x' can be found by minimizing:

$$(M_i(x') - T)^2 \tag{3}$$

where T is a target value smaller than zero. It is well-known [44] that minimizing a loss of the form:

$$(M_i(x') - T)^2 + \lambda ||x' - x||_2^2 \tag{4}$$

is equivalent to finding a minimizer of Eq. (3) that lies in the ball defined by $||x - x'||_2^2 \leq \rho$ for some ρ. As such, minimizing this objective for an appropriate value of λ and T is a good strategy for finding adversarial perturbations of image x with small ℓ_2 norm.

Writing $C^{(l)}(x)$ for the classifier response of the l^{th} layer of the neural net, we consider the related loss:

$$(M_i(x') - T)^2 + \lambda' \sum_{l \in \mathcal{L}} ||C^{(l)}(x') - C^{(l)}(x)||_2^2 + \lambda ||x' - x||_2^2 \tag{5}$$

defined over a set of layers of the neural network $\mathcal{L}$.

The second term is the perceptual loss of [20], and minimizing this objective is equivalent to finding a minimizer of Eq. (4) subject to the requirement that x' lies in the ball defined by $\sum_{l \in \mathcal{L}} ||C^{(l)}(x') - C^{(l)}(x)||_2^2 \leq \rho'$ for some ρ'.

To convert the adversarial perturbation to a saliency map, we first calculate the size of the adversarial perturbation in each pixel by computing the average squared difference over the channels. Second, in order to highlight areas with large changes, we apply a Gaussian blur with parameter σ to the differences to give our resultant saliency map.

We systematically evaluate the effect of altering the regularized layers for a range of tasks. We find that the method is relatively stable and Eq. (5) performs better than the unregularized Eq. (4) for weak localization, insertion and deletion and the pointing game. As shown in Fig. 2, as more layers are regularized the perturbation becomes more localized.

4. Perceptual Perturbations as Explanations

Before describing our experimental overview, we give a qualitative analysis of the perceptual perturbations (Fig. 4). The perturbations do a good job of localizing on a single object class, even in the presence of highly textured images (dragonfly on fern), and in images with multiple classes (baseball and people). Some error in localization seems to arise from supporting classes adjacent to the object - e.g. human legs behind the lawnmower are found to be salient.

Furthermore, a qualitative evaluation can be seen in Fig. 3. These images were selected to be challenging – we visualize a subset of those images where unregularized adversarial perturbations did not align with the object. Compared to other visual explanation techniques, our method highlights the interior textures of the target object in the image. This differs from gradient-based methods which capture finer edge details such as SmoothGrad [42] and to activation-based methods which highlight the entire object coarsely such as Grad-CAM [38]. This is perhaps clearest in the first image where we capture the interior texture of the socks rather than just its hard contours.

4.1. Sensitivity Studies

For all experiments we performed extensive sensitivity studies to evaluate the importance of regularizing over different layers (Fig. 5). Certain trends can be detected. Regularizing over most of the layers is effective for weak localization, insertion and deletion, and pointing games. However, the pointing game performs best with different layers, possibly as you only need to find a single point of an object, we find that regularizing only the top layer is optimal in our ablation study, even when testing at multiple resolutions.

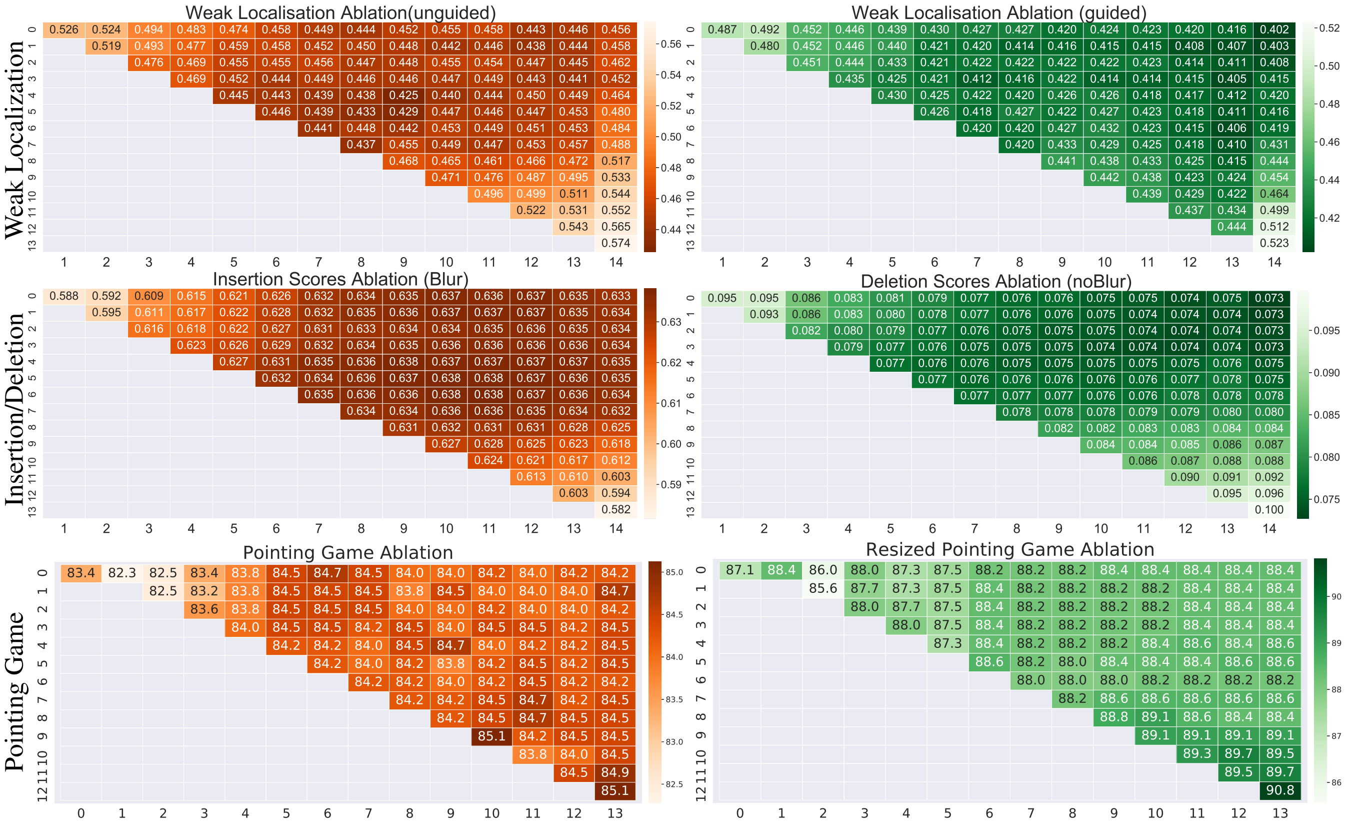

Figure 5. Sensitivity study on the choice of layers regularized for the three games where the ith and jth entry displays the performance from regularizing layers $\{i, i+1, \ldots, j-1\}$. *Top* shows the performance of weak localization for the unguided variant and the guided variant (Lower is better). The best performance is achieved when we regularize ReLU layers from 4 to 9 for the unguided variant and from 0 to 14 for the guided variant. *Middle* shows the performance of deletion for the unBlur variant (lower is better) and insertion for the blur variant (higher is better). ReLU layers from 0 to 12 have been selected as it achieves the largest deviation between deletion score and insertion score. *Lower* shows pointing game, with left panel representing orig. image size and right the resized image size. We regularize by each set of ReLU layers from ReLU 0 to ReLU 12, with $i = j = 0$ equivalent to the no perceptual approach, and report the average success rate (higher is better). For ease of presentation in this game we present the best performance over $\sigma \in \{0, 1, \ldots 100\}$.

4.2. Weak Localization

We evaluate perceptual perturbations as explanations using the weak localization protocol [12], and test our approach on the first 2000 ImageNet [37] validation images.

We construct a set of bounding boxes for the largest connected region using three simple strategies based on: thresholding the raw values, thresholding a fixed percent of the image, and thresholding scaled by the image mean following [12]. To match the previous thresholding strategies we normalize individual saliency maps to be in the range of $[0, 1]$ before applying the blur. For the first strategy, we use a value threshold where we grid search over the set of thresholds α where $0 < \alpha \leq 1$ at intervals of size 0.05. For the second strategy, we use a percentage threshold where we consider the $\alpha\%$ most salient pixels grid search over the same interval. For the third strategy, we use a threshold scaled by the per image mean where we grid search over the set of thresholds α where $0 < \alpha \leq 10.50$ at intervals of size 0.05. We report the scores on all three strategies as well as the optimal strategy for each explanation method. For each threshold, we extract the largest connected component and draw a bounding box around it. The object is considered to be successfully localized when the Intersection over Union measure between this box and the ground truth is at least 0.5. Following Grad-CAM's guided version [38], which makes use of image gradients, we consider a guided variant of our own method consisting of an element-wise multiplication between our perturbations and the normalized gradient of the $C_i(x)$ with respect to the image x.

We set $T = -2$, $\lambda' = 10000, \lambda = 1$ in Eq. (5), and we run an ablation study for the first 1000 images. We select two sequential sets of ReLU layers to regularize over in a VGG19bn network [41] using the value-threshold strategy. For the un-guided variant of our method, we regularize from

Method	Value	Per.	Mean	Best
GuidedBP [45]	0.48	0.52	0.49	0.48
Grad-CAM [38]	0.49	0.51	0.48	0.48
Guided-Grad-CAM [38]	*0.46*	0.49	*0.45*	0.45
Excitation [56]	0.48	0.50	0.44	*0.44*
SmoothG [42]	0.47	0.49	0.47	0.47
IntegratedG [47]	0.44	0.51	0.48	*0.44*
Extremal [10]	0.55	0.52	0.54	0.52
RISE [33]	0.51	0.52	0.48	0.48
NormGrad [34]	0.49	0.52	0.46	0.47
sNormGrad [34]	0.49	0.52	0.47	0.47
Us NoPer	0.50	*0.47*	0.46	0.46
Us Unguided	0.44	0.44	**0.43**	0.43
Us Guided	**0.41**	**0.43**	**0.43**	**0.41**

Table 1. Results for Weak Object Localization (lower is better, see sec.4). We have the lowest error for each thresholding strategy.

Method	Deletion Score	Insertion Score
Gradient [40]	0.19	0.51
Deconv [55]	0.21	0.56
GuidedBP [45]	0.14	0.57
Excitation [56]	0.12	*0.63*
Grad-CAM [38]	0.11	0.64
Extremal [10]	0.16	0.62
RISE [33]	0.12	**0.65**
NormGrad [34]	0.09	0.58
sNormGrad [34]	*0.10*	0.59
blurDiff ($\sigma = 2.5$)	0.14	0.59
Us NoPer ($\sigma = 0.0$)	*0.10*	0.42
Us NoPer ($\sigma = 2.5$)	0.15	0.54
Us ($\sigma = 0.0$)	**0.07**	0.54
Us ($\sigma = 1.0$)	0.09	0.61
Us ($\sigma = 2.5$)	0.11	0.62
Us ($\sigma = 5.0$)	0.12	*0.63*

Table 2. Results for deletion (first column) and insertion game (second column). The result shows our method performs better than other methods for the deletion metric (without blur) and is comparable to other methods for the insertion metric (with blur).

ReLU 4 to 9 (0 indexed) and for the guided variant we regularize from ReLU 0 to 14. We report the results for each strategy using these layers. Qualitative evidence suggests that regularizing sequentially more and higher layers tends to improve object localization in the image (see Figs 2,5).

We compare our method, its guided variant and our method without the perceptual loss to other methods [10, 33, 34, 38, 42, 45, 47, 56]. For the methods in [38, 42, 45, 47], we used the PyTorch CNN Visualizations repository [28]. For [10, 56], we use the TorchRay implementation [9] and for [34], we use its NormGrad branch [11]. For RISE [33] we used the authors code from [32]. We used the provided parameters for all methods.

We varied $\sigma \in [1, 30]$ and found that results saturate at $\sigma = 20$ for all three thresholds. We outperform all other methods (see Table 1). Our guided variant obtains the lowest error for two thresholding strategies and joint best using mean-based thresholding. The non-guided version performs second-best in value thresholding, percent-based thresholding, and joint best in the mean-based thresholding strategy. When selecting the best possible threshold for each method, our method and it's unguided variant achieve the lowest and second lowest error rates respectively.

4.3. Insertion/Deletion Game

We compute the insertion and deletion metrics from [33]. For the deletion metric, we construct the deletion response curve by sequentially changing the most salient pixels from their original value to mid-gray and measuring the classifier response. The deletion metric is defined as the AUC of the deletion curve, a smaller AUC score (i.e. a sharper drop in classifier response) is considered indicative of a better explanation. The insertion metric is similar, however, rather than removing the pixels of largest saliency, it inserts the original values into a blurred version of the original image. The metric is again the AUC of the insertion curve, however a higher AUC score is considered better as this corresponds to a sharper increase in classifier response with the addition of the most salient pixels.

We set $T = -2, \lambda' = 10000, \lambda = 1$ in Eq. (5), and we run an ablation study for the first 500 images. We identify the ReLU layers from 0 to 12 to regularize over in a VGG19bn network that achieve an optimal difference between having a high insertion score and a low deletion score (see Fig. 5). We found that Gaussian blurring the saliency map improves the performance for the insertion metric, while it decreases the performance of the deletion metric. We varied the blur width between 1 and 30 and found the performance for insertion saturates when $\sigma \geq 4.0$. For consistency, we used the blur function in [33].

We compare our perceptual method and its blurred variants with a set of alternative methods for this experiment [10, 33, 34, 38, 40, 45, 55, 56]. We use the standard implementation of this game from the RISE repository [33, 32]. We compare our method with RISE from the above package and each of the saliency methods implemented in the TorchRay package [9] using the standard settings. For NormGrad we used its branch in the TorchRay Repository [11]. We include two additional baselines. The first baseline is our method without the perceptual loss and the second baseline is the ℓ_2 norm between the image and a blurred version. Our approach performs significantly better than our baseline and is noticeably better than all others for the deletion metric and joint third for the insertion metric (see Table 2).

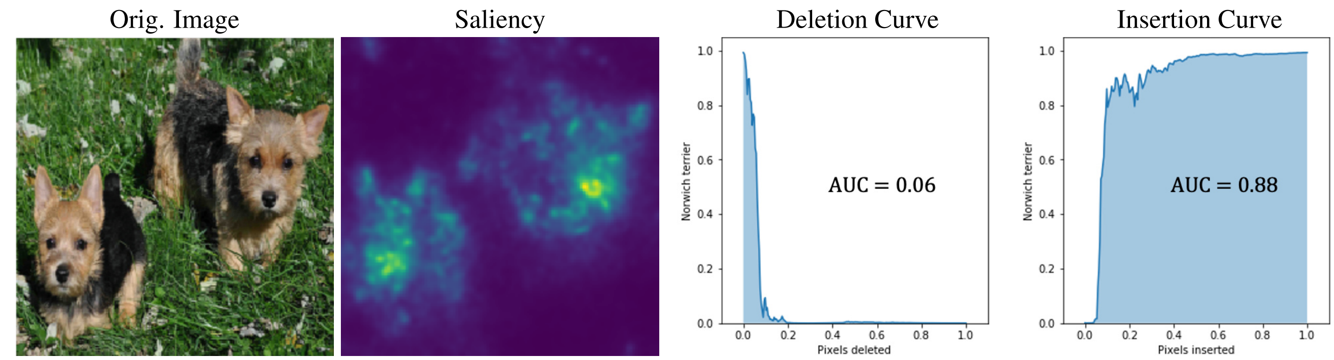

Figure 6. Insertion (higher is better) and Deletion (lower is better) metrics [33]. The average AUC over 5000 images in the ImageNet validation set as pixels are sequentially set to, or changed from, their value in the original image.

4.4. Pointing Game

Additionally, we evaluated on the pointing game introduced by [56] on the VOC dataset. In this game the task is to select a single point in the image which is in the object in question [56]. The game is subdivided into 2 subtasks, a standard set of images, and a more difficult subset in which the object occupies less than 25% of the image, and must also contain a distracting class (see [56]). We use the TorchRay [9] implementation of the game and comparison methods and [11] for the implementation of NormGrad.

The pointing game uses a modified VGG16 classifier that assigns multiple classes to a single region of the image, and thus we use Eq. (5) with $T = -10$, $\lambda' = 1000$, $\lambda = 1$ and the formulation from Eq. (2). In this game, the images vary in size, and the classifier returns a detector like response for a set of overlapping regions of the image. Given a choice of class, we suppress all candidate regions by ensuring that the maximal response of any region is close to the target value.

As with the previous games, we select the layers and blur using an ablation study. We test each range of contiguous layers from the first ReLU layer, to final ReLU layer, and we vary the blur between 0 and 100[4]. We perform the study on the first 500 bounding boxes, which corresponds to 457 standard bounding boxes of which 160 are in the difficult subset and 43 boxes excluded by the benchmark.

Results can be seen in Fig. 5 lower left. Due to the reduced dataset (457), we report the success rate over the standard set, rather than the average class success reported on the full dataset. For compactness, we display the best result over the σs, and to measure consistency we present the percentage of σs which yield a performance above 82% (SM B.2). Two layer sets achieve the largest score, ReLU 9 (layer 22) with $\sigma \in \{47, 48\}$ or ReLU 12, with $\sigma \in \{41, 42\}$. We select ReLU 9 with $\sigma = 48$. The first layer is less sensitive to the choice of blur with $\approx 70\%$ giving a result above 82% in contrast to the $\approx 44.6\%$ for the final layer (this setting also gives a lower performance overall with a score of 83.8% (66.4%)).

[4]For efficiency, we create one perturbation per set of layers and vary σ

Method	Orig. Image	Scaled Image
Center	69.6 (42.4)	69.6 (42.4)
Gradient [40]	76.3 (56.9)	84.6 (70.0)
Deconv [55]	67.5 (44.2)	75.0 (53.5)
GuidedBP [45]	75.9 (53.0)	83.7 (67.1)
Excitation [56]	77.1 (56.6)	84.0 (67.5)
Grad-CAM [38]	*86.6 (74.0)*	89.1 (77.7)
Extremal [10]	**88.0 (77.3)**	86.4 (71.0)
RISE [33]	86.7 (75.4)	NA (NA)
NormGrad [34]	81.9 (64.9)	*88.6* (75.6)
sNormGrad [34]	86.0 (72.7)	**90.1 (80.8)**
Us NoPer	81.2 (62.9)	86.0 (71.6)
Us	85.1 (69.0)	88.2 (*76.5*)

Table 3. Results for the Pointing Game. We present two sets of results, the performance on the images in the original size (first column), and second the performance on images which are scaled to 1.5x using bilinear scaling (second column). For each result, the first number shows the percentage on the standard set, with the number in brackets the performance on the difficult set.

The full results, conducted on all images/tests are presented in the left column of Table 3. We are competitive with the best methods, with a 2.9% performance difference with the best method in the standard setting. Further, in SM B we present a study with a limited set of σs where we achieve qualitatively similar results, but only a 0.3% drop in standard performance and no drop in the difficult setting. We additionally compare to our framework with the same parameters but without the perceptual loss (with the blur which maximizes the ablation score - see SM B.1). We refer to as this 'no perceptual'. As we see in Table 3 our version with the perceptual loss outperforms it. We also display this approach in the ablation study (top left corner). The perceptual loss outperforms in almost all cases, with exceptions in the low layers. Further, the results with the perceptual are more consistently above 82% (SM B).

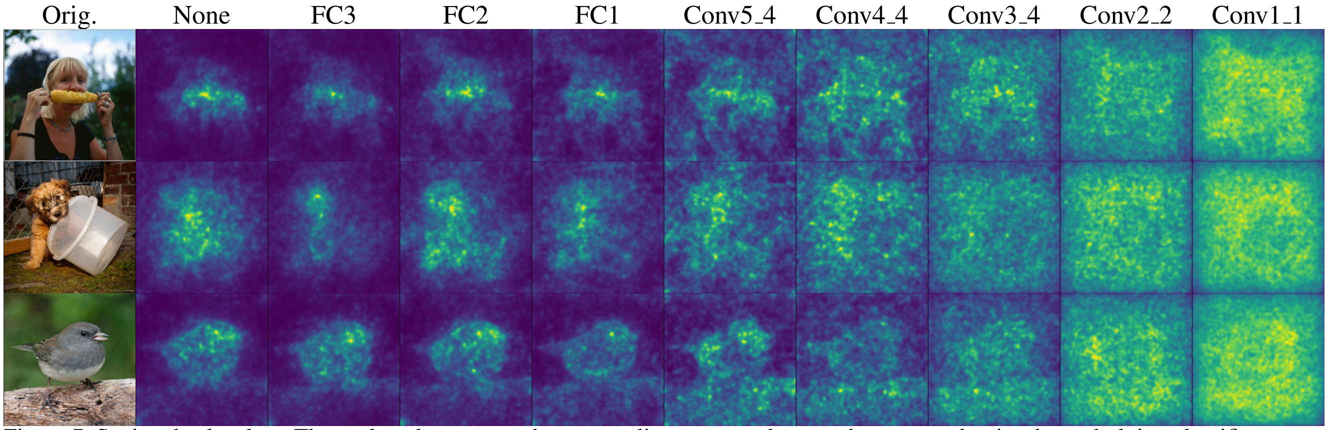

Figure 7. Sanity checks plots. These plots demonstrate how our saliency maps change when we randomise the underlying classifier.

Finally, we also propose an additional pointing game experiment. As this game is defined on non-standard sized image, we consider the case where we increase the size of the images via a simple resizing, and test the performance on this set. This helps performance as the fully convolutions network then gives predictions for more areas that are then maximized over. We implement this as a pre-and post-processing step, where we resize the image, construct the saliency map, and resize the saliency map to the correct size. We perform a similar ablation study for this case, (Fig. 5 bottom right panel). The optimal layer in this case is ReLU 12 with $\sigma \in \{28.0, 30.0\}$, (selected 30).

The results can be seen in the right column of Table 3, where most methods work on this new dataset (the TorchRay RISE implementation requires a perfectly sized image and is NA). The performance of many of the methods is increased in this new approach, with many weaker methods substantially increasing (except for center - a baseline which chooses the center point). Further, the performance characteristics of the best performing methods also improves with resizing, albeit by lower margins, indicating that this resizing generally helps almost all methods.The best performing method is Selective NormGrad with a score of 90.1%, 2.1% above the previously highest score. Our Perceptual method is competitive with the best performing methods, achieving a score of 88.2%.

4.5. Sanity Checks

Finally, we apply the Sanity Checks proposed in [1] to our method. We randomize the weights on the final k layers of the network, and observe how the saliency of our approach varies visually. We perform this experiment on VGG19bn, as most of the experiments in this paper were performed on this network, and we match the remaining parameters/layers to the insertion deletion study. When randomizing the layers we set the parameters to the value that would have been in an untrained network. We select the layers in the VGG19bn architecture that correspond to the VGG16 layers used by [34], namely each fully connected layer and the final convolution in each set with the exception of the first set for which we use the first convolution.

Fig. 7 shows the results on three common sanity check images. As the layers are progressively randomized, saliency spreads from the relevant objects towards other objects and highly textured regions in the image, and is mostly non-existent after the second convolution qualitatively matching the behavior in Selective NormGrad seen in [34].

5. Conclusion

We explored a novel regularization for adversarial perturbations based on the perceptual loss. This regularization is designed to block the exploitation of exploding gradients when generating adversarial perturbations forcing larger and more meaningful perturbations to be generated. The fact that they remain imperceptible to humans is another piece of the puzzle in understanding the interrelationship between adversarial perturbations, neural networks, and human vision. We believe that the imperceptible nature of our adversarial perturbations is due to both explanations discussed in the introduction for the existence of adversarial perturbations being partially correct. Even when regularizing over the layers of our network and preventing adversarial perturbations from exploiting exploding gradients and some of the inherent instability in deep networks, Szegedy et al.'s [48] argument still holds and one can obtain a new class label by slightly altering a large number of pixels.

We have shown how these perturbations can be interpreted as explanations and obtained state-of-the-art results on several standard explainability benchmarks.

Acknowledgments This work was supported by the Omidya Group and The Alan Turing Institute under the UK Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/N510129/1 and Accenture Plc. Moreover, we acknowledge Pearl for computing resources and in particular the help of Tomas Lazauskas and Suleman Tariq.

References

- [1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 9505–9515. Curran Associates, Inc., 2018.
- [2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. volume 80 of *Proceedings of Machine Learning Research*, pages 284–293, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.
- [3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.
- [4] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *ICLR*, 2019.
- [5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar 2018.
- [6] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *Thirty-second AAAI conference on artificial intelligence*, 2018.
- [7] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *NeurIPS*, pages 592–603, 2018.
- [8] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems*, pages 3910–3920, 2018.
- [9] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Torchray. https://github.com/facebookresearch/TorchRay, 2019.
- [10] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [11] Ruth Fong, Sylvestre-Alvise Rebuffi, Xu Ji, and Andrea Vedaldi. Torchray. github.com/ruthcfong/TorchRay/tree/normgrad, 2019.
- [12] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *ICCV*, Oct 2017.
- [13] Jacob R Gardner, Paul Upchurch, Matt J Kusner, Yixuan Li, Kilian Q Weinberger, Kavita Bala, and John E Hopcroft. Deep manifold traversal: Changing labels with convolutional features. *arXiv:1511.06421*, 2015.
- [14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2019.
- [15] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv:1801.02774*, 2018.
- [16] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.
- [17] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384, Long Beach, California, USA, 09–15 Jun 2019. PMLR.
- [18] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Generating counterfactual explanations with natural language. *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning*, pages 95–98, 2018.
- [19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.
- [20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.
- [21] Simran Kaur, Jeremy Cohen, and Zachary C. Lipton. Are perceptually-aligned gradients a general property of robust classifiers? *Science Meets Engineering of Deep Learning" Workshop at NeurIPS 2019*, 2019.
- [22] Ann B. Lee, Kim S. Pedersen, and David Mumford. The nonlinear statistics of high-contrast patches in natural images. *International Journal of Computer Vision*, 54(1):83–103, 2003.
- [23] David Lewis. *Counterfactuals*. John Wiley & Sons, 1973.
- [24] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *ICLR*, 2017.
- [25] Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Sparsefool: a few pixels make a big difference. In *CVPR*, pages 9087–9096, 2019.
- [26] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. *CVPR*, Jun 2016.
- [27] A. Mustafa, S. Khan, M. Hayat, R. Goecke, J. Shen, and L. Shao. Adversarial defense by restricting the hidden space of deep neural networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3384–3393, 2019.
- [28] Utku Ozbulak. Pytorch cnn visualizations. https://github.com/utkuozbulak/pytorch-cnn-visualizations, 2019.
- [29] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

[30] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2, 2012.

[31] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

[32] Vitali Petsiuk. Rise: Randomized input sampling for explanation of black-box models. https://github.com/eclique/RISE, 2018.

[33] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[34] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[35] M Ribeiro, S Singh, and C Guestrin. "why should i trust you?"explaining the predictions of any classifier. *SigKDD*, 2016.

[36] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. The odds are odd: A statistical test for detecting adversarial examples. In *ICML*, pages 5498–5507, 2019.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[38] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[39] Carl-Johann Simon-Gabriel, Yann Ollivier, Leon Bottou, Bernhard Schölkopf, and David Lopez-Paz. First-order adversarial vulnerability of neural networks and input dimension. volume 97 of *Proceedings of Machine Learning Research*, pages 5809–5817, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[40] K Simonyan, A Vedaldi, and A Zisserman. Deep inside convolutional networks. *ICLR*, 2013.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

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

[43] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018.

[44] Danny C Sorensen. Newton's method with a model trust region modification. *SIAM Journal on Numerical Analysis*, 19(2):409–426, 1982.

[45] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv:1412.6806*, 2014.

[46] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6976–6987, 2019.

[47] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv:1703.01365*, 2017.

[48] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[49] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1633–1645. Curran Associates, Inc., 2020.

[50] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[51] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. *arXiv:1907.02584*, 2019.

[52] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

[53] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gpdr. *Harv. JL & Tech.*, 31:841, 2017.

[54] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[55] M Zeiler and R Fergus. Visualizing and understanding convolutional networks. *ECCV*, 2014.

[56] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *Int. J. Comput. Vis.*, 126(10):1084–1102, Dec 2017.

[57] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CVPR*, Jun 2016.