International Journal of Multimedia Information Retrieval (2021) 10:127–140
https://doi.org/10.1007/s13735-021-00208-3

REGULAR PAPER

Counterfactual attribute-based visual explanations for classification

Sadaf Gulshad[1] · Arnold Smeulders[1]

Received: 1 February 2021 / Revised: 2 March 2021 / Accepted: 8 March 2021 / Published online: 18 April 2021

Abstract
In this paper, our aim is to provide human understandable intuitive factual and counterfactual explanations for the decisions of neural networks. Humans tend to reinforce their decisions by providing attributes and counterattributes. Hence, in this work, we utilize attributes as well as examples to provide explanations. In order to provide counterexplanations we make use of directed perturbations to arrive at the counterclass attribute values in doing so, we explain what is present and what is absent in the original image. We evaluate our method when images are misclassified into closer counterclasses as well as when misclassified into completely different counterclasses. We conducted experiments on both finegrained as well as coarsegrained datasets. We verified our attribute-based explanations method both quantitatively and qualitatively and showed that attributes provide discriminating and human understandable explanations for both standard as well as robust networks.

Keywords Explainable AI · Counterfactual · Explanations · Attributes · Classification · Adversarial examples

1 Introduction

When deploying machine learning and computer vision models in the real world it is of utmost importance that we explain the decisions made by these models in human understandable and intuitive way. The preferable procedure to provide such explanations would be as humans explain their decisions. For example, when a person classifies a bird into the "***Cardinal***" class the reason provided by the person is because it has a "***Crested head***" and a "***Red beak***". Humans also tend to support their decisions by providing counterexamples and counterattributes such as, this bird would be classified into the class "***Pine Grosbeak***" if it will have a "***Plain head***" and a "***Black beak***" as shown in Fig. 1. Inspired by human explanations, in this paper we employ human understandable visual attributes for providing factual and counterfactual explanations.

A large body of work in explainable AI focuses on explaining the decisions of neural network-based classifiers using saliency maps [35,42]. Saliency maps highlight the part of the image which supports the classification however, the support to the classification might be distributed across the whole image, or might lie in the color or texture of the object. Hence, it becomes difficult to localize the part of the image responsible for the classification especially for fine-grained datasets. Furthermore, saliency maps tell us about what is present in the image and do not provide any information about what is absent i.e counterfactual information. Therefore in this work, we focus on human nameable attributes for providing the reasoning behind specific decisions and perturbations to arrive at attributes belonging to counterclasses to provide counterfactual explanations.

In a closely related work [13], the authors provided counterfactual explanations for classification decisions by replacing the part of the original image with the similar part from the distractor image belonging to the counterclass, such that the class of the image changes. However, their method is pixel-based, hence requires matching imaging conditions such as pose and illumination. In contrast in this work, we introduce perturbations in the images so that the attribute values change to the counterclass attribute values.

In a recent work for the different purpose of enhancing the generalization power of visual question answering systems [1] authors utilized counterfactuals and trained the network with counterexamples. Similarly, in our work, we improve the generalization and robustness of the neural network-based classifier by training it with counterexamples.

✉ Sadaf Gulshad
s.gulshad@uva.nl

Arnold Smeulders
a.w.m.smeulders@uva.nl

[1] UvA-Bosch Delta Lab, University of Amsterdam, Amsterdam, The Netherlands

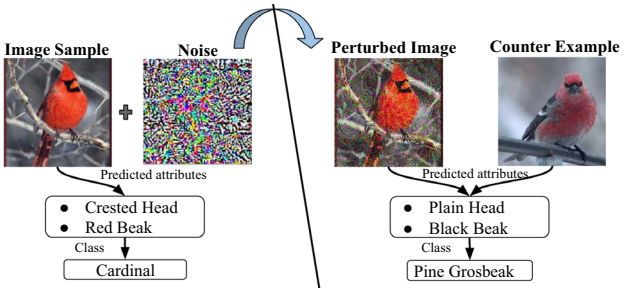

Fig. 1 We use attributes to explain why an image on the left is classified into the *Cardinal* class rather than the *Pine Grosbeak* class on the right. And we use attributes with examples to explain when it will be classified as a *Pine Grosbeak* by exploiting perturbed examples and their attribute values. We show that when the predicted attributes for the image change from "*Crested Head*" and "*Red Beak*" to "*Plain Head*" and "*Black Beak*", the image will be classified as *Pine Grosbeak* (color figure online)

However, we go a step further and provide counterfactual explanations for this network too.

We define the closeness of classes in the embedding space based on the attribute similarity and evaluate our method when images get misclassified into the closer counterclass [37] as well as when we force them to be misclassified into a distant counterclass [6]. We complement our attribute-based explanations with counterexample-based explanations by selecting the examples containing counterattributes.

Our main contributions are given as follows:

- We provide novel explanations for classification decisions by utilizing intuitive factual and counterfactual attributes and examples.
- We study the change in attribute values when images are perturbed to provide counterfactual explanations from any alternative counterclass as well as when images are perturbed to provide counterfactual explanations from our desired counterclass.
- We propose a novel method to assist our attribute-based explanations with counterexamples, selected based on these counterattributes.

We evaluate our attribute-based explanations *quantitatively* and *qualitatively* for a *standard* as well a *robust* network. Our results on three different datasets of varying sizes and granularity show that attributes provide effective factual and counterfactual explanations for classifier decisions. This paper is an extended version of our conference paper [14].

2 Related work

Explaining the output of a decision maker is commonly motivated by the need to build user trust before deploying them into a real world environment [11,15,26].

2.1 Explainability

Previous work for visual classification explanation is broadly grouped into two types: (1) *rationalization*, that is, justifying the network's behavior and (2) *introspective explanation*, that is, showing the causal relationship between input and the specific output [10]. The first group has the benefit of being human understandable, but it lacks a causal relationship between input and output. The second group incorporates the internal behavior of the network, but lacks human understandability. In this work, we explain the decisions of neural networks in the human style of explanations by singling out specific attributes for positive evidence when the image is classified correctly and by following specific attributes for negative evidence when the image is directed for misclassification in a counterclass.

An important group of work on understandibility focuses on text-based class discriminative explanations [16,30], text-based interpretation with semantic information [9] and generating counterfactual explanations with natural language [17], they all fall in the *rationalization* category. Text-based explanations are orthogonal to our attribute-based explanations as attributes tend to deliver the key-words in the sentence and carry the quintessence for the semantic distinction. Particularly for fine-grained classification all sentences for all classes tend to display the same structure hence, the core of the semantic distinction between classes lies in attributes where we put our emphasis. Generating sentences is valuable but largely orthogonal to our approach.

To tackle the similar task of explaining visual decisions, there is the large body of work on activation maximization [35,42], learning the perturbation mask [12], learning a model locally around its prediction, and finding important features by propagating activation differences [32,34]. They all fall in the group of *introspective explanations*. All these approaches use saliency maps for explanation. We observe that saliency maps [33] are frequently weak in justifying classification decisions, especially for fine-grained images. For instance, in Fig. 2 the saliency map of a clean image classified into the ground truth class, "red-winged blackbird", and the saliency map of a misclassified perturbed image, look quite similar. Instead, by grounding the predicted attributes, one may infer that the "orange wing" is important for "red-winged blackbird" while the "red head" is important for "red-faced cormorant". Indeed, when the attribute value for orange wing decreases and for red head increases the image gets misclassified. Therefore, we propose to predict and ground attributes for both clean and perturbed images to provide visual as well as attribute-based interpretations.

Counterfactual explanations Explanations which consider counterdecisions or counteroutcomes are known as *counterfactual explanations* [25]. An interesting approach in a recent paper [13] proposes to generate counterfactual expla-

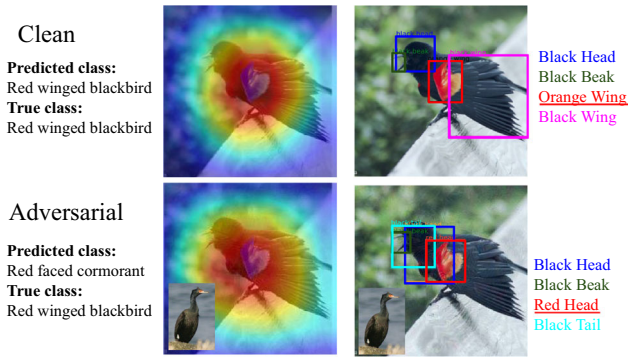

Fig. 2 Fine-grained images are difficult to explain with saliency maps: when the answer is wrong, often saliency-based methods (left) fail to detect what went wrong. Instead, attributes (right) provide intuitive and effective visual and textual explanations

nations by selecting a distractor image from a counterclass and replacing the region in the input image with a region from the distractor image such that the class of the input image changes into the class of the distractor image. Pixel-based replacements pose high restrictions on the similarity of viewpoint, pose and scene between the two images, which makes the selection and replacement of the patches difficult. We follow the same inspiration of human-motivated counterexamples. However, our approach focuses on attributes for generating explanations, as they contain the semantic core of the distinction between two competing classes and, attributes can naturally incorporate large changes in imaging conditions of size, illumination and viewpoint. Additionally, we use perturbations to change the class of the input image we analyze which attributes lead to the change in class.

Another closely related work, [21], focuses on the multimodal complimentarity of text and image for explanations. They maximize the interaction information between class predictor and explanation generator by simultaneously training them using variational lower bound. However, by the nature of their method their example-based explanations will be visually completely different from the input image. In our work, by using the method of directed perturbations and discriminating attributes we are capable of selecting the most critical counterexamples as the most effective explanations.

In [1], authors utilized counterfactuals for enhancing the generalization and applicability of visual question answering systems. However, in our work for providing explanations we increase the generalization and robustness of neural network classifier by training it on counterfactuals. After robustification we verify our method on the robustified network by studying the change in attributes for it.

2.2 Adversarial examples

Untargeted methods Small, carefully crafted perturbations, called *adversarial perturbations*, have been used to alter the inputs of deep neural networks, which results in *adversarial examples*. These adversarial examples drive the classifiers to the wrong class [37]. Such methods of directed perturbations include iterative fast gradient sign method (IFGSM) [23], the Jacobian-based saliency map attacks [29], one pixel attacks [36], Carlini and Wagner attacks [7] and universal attacks [28]. Here, our aim is to utilize the directed noise from adversarial examples to study the change in attribute values. Therefore, we select the IFGSM-method which is fast and strong for our experiments to lead images into counterclasses.

Targeted methods When small adversarial perturbations are introduced in the images to misclassify them into the desired counter classes are called *targeted attacks* [8]. Targeted attacks are stronger and more difficult to achieve than *untargeted attacks* because the algorithm needs to find the perturbations, which will misclassify the image into the desired class instead of misclassification into any alternative class [6] i.e *untargeted attacks*. Besides studying the change in attribute values for untargeted attacks, here we also study the change in the attribute values when images are directed into desired classes. For this purpose we utilize the targeted version of IFGSM method and compare the results for untargeted and targeted attacks to verify whether our proposed attribute-based counterfactual explanations also function for targeted attacks.

2.3 Adversarial examples for explainability

Adversarial examples have been used for understanding neural networks. [18] aims at utilizing adversarial examples for understanding deep neural networks by extracting the features which provide the support for classification into the target class. In this paper instead of providing feature based visualizations we focus on human understandable attributes for providing explanations for decisions. In [20], the authors proposed a data-path visualization module consisting of the layer level, the feature level, and the neuronal level visualizations of the network for clean as well as for adversarial images. In contrast, we focus on exploiting adversarial examples to generate intuitive factual and counterfactual human understandable explanations with attributes and visual examples.

In [40], the authors investigated adversarially trained robust convolutional neural networks by constructing input images with different textual transformations while at the same time preserving the shape information. They do this to verify the shape bias in adversarially trained networks compared with standard networks. Similarly, in [38], the authors

showed that saliency maps from adversarially trained robust networks align well with human perception. In our work, we also provide explanations when an image is correctly classified with an adversarially trained robust network and verify that the attributes predicted by our method with a robust network still retain their discriminative power for explanations.

Adversarial examples and counterfactual explanations In a closely related work [19] authors reveal the duality relationship between adversarial examples and explanations. They argue that adversarial examples could be generated from counterexamples and counterexamples could be generated from adversarial examples. We follow a similar idea but instead propose to utilize adversarial examples for explanations in the presence of human understandable attributes.

Similarly, [5] tries to solve the paradox that previous research [39] shows that adversarial examples and counterfactual explanations are equivalent, then where lies the difference between them? They argue that this paradox could be solved by properly studying the semantics (i.e., neuronal activations) of counterfactuals for providing explanations. In this paper instead of focusing on solving the paradox between adversarial examples and counterfactual explanations we make use of adversarial examples with attributes to provide counterfactual explanations.

3 Method

3.1 Adversarial perturbations

Given n-th image x_n and its respective ground truth class y_n predicted by a classifier $f(x_n)$, an image $\hat{x}_n$ is generated by adding adversarial perturbations to it such that the classifier $f(\hat{x}_n)$ predicts y, where $y_n \neq y$, and x_n and $\hat{x}_n$ are close according to some distance metric. Next, we present the method for generating adversarial examples through *untargeted attacks* [6] and *targeted attacks* [6,8].

Untargeted attacks We leverage IFGSM method [23] to generate adversarial perturbations. The mechanism for generating adversarial examples through basic iterative method is given by:

$$\hat{x}_n^0 = x_n$$
$$\hat{x}_n^{i+1} = \text{Clip}_\epsilon\{\hat{x}_n^i + \alpha \text{Sign}(\nabla_{\hat{x}_n^i}\mathcal{L}(\hat{x}_n^i, y_n))\} \quad (1)$$

where, $\hat{x}_n^0$ is the input image at step $i = 0$, $\nabla_{\hat{x}_n^i}\mathcal{L}$ is the derivative of the loss function w.r.t to the current input image, α is the step size taken at step i in the direction of sign of the gradient, and finally the result is clipped by Clip_ϵ.

Targeted attacks For targeted attacks we target our input image to be misclassified into a specific class y_t. Following equations are used to create adversarial perturbations for misclassification in to a target class.

$$\hat{x}_n^0 = x_n$$
$$\hat{x}_n^{i+1} = \text{Clip}_\epsilon\{\hat{x}_n^i - \alpha \text{Sign}(\nabla_{\hat{x}_n^i}\mathcal{L}(\hat{x}_n^i, y_t))\} \quad (2)$$

In the targeted attacks we maximize the loss against ground truth class y_n and minimize the loss against target class y_t.

3.2 Adversarial robustness

Adversarial training Adversarial training [37] is one of the state of the art method for robustness against adversarial perturbations. In adversarial training the model $f^r(\hat{x}_n)$ finds the worst case adversarial examples and trains the network on these adversarial examples besides training it on clean images to make it robust against adversarial perturbations. Hence, this leads to an improvement in performance against adversarial perturbations. The following objective function is minimized in adversarial training:

$$\mathcal{L}_{adv}(x_n, y_n) = \gamma\mathcal{L}(x_n, y_n) + (1-\gamma)\mathcal{L}(\hat{x}_n, y) \quad (3)$$

where, $\mathcal{L}(x_n, y_n)$ is the classification loss for clean images, $\mathcal{L}(\hat{x}_n, y)$ is the loss for adversarial images and γ regulates the loss to be minimized.

3.3 Attribute prediction

We use class attributes available with the dataset to predict per image attributes and provide explanations for classification. The model is shown in Fig. 3. At training time our network learns to map image features closer to their ground truth class attributes and farther from other classes in the embedding space. During test time when clean image features are projected in the learned embedding space the image gets mapped closer to the ground truth class attributes e.g. "Crested head" and "Red beak" associated with the ground truth class "Cardinal", see Fig. 3. However, an adversarially perturbed image gets mapped closer to the wrong class attributes e.g. "Plain head" and "Black beak" belonging to the counterclass 'Pine Grosbeak", Fig. 3.

Given the n-th input image features $\theta(x_n) \in \mathcal{X}$ and output class attributes $\phi(y_n) \in \mathcal{Y}$ from the sample set $\mathcal{S} = \{\theta(x_n), \phi(y_n), n = 1 \ldots .N\}$ we employ SJE [2] to predict attributes in an image. SJE learns to map $\mho : \mathcal{X} \longrightarrow \mathcal{Y}$ by minimizing the empirical risk of the form $\frac{1}{N}\sum_{n=1}^{N}\Delta(y_n, \mho(x_n))$, where $\Delta : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ estimates the cost of predicting $\mho(x_n)$ when the ground truth label is y_n.

A compatibility function $F : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is defined between input $\mathcal{X}$ and output $\mathcal{Y}$ space:

$$F(x_n, y_n; W) = \theta(x_n)^T W \phi(y_n) \quad (4)$$

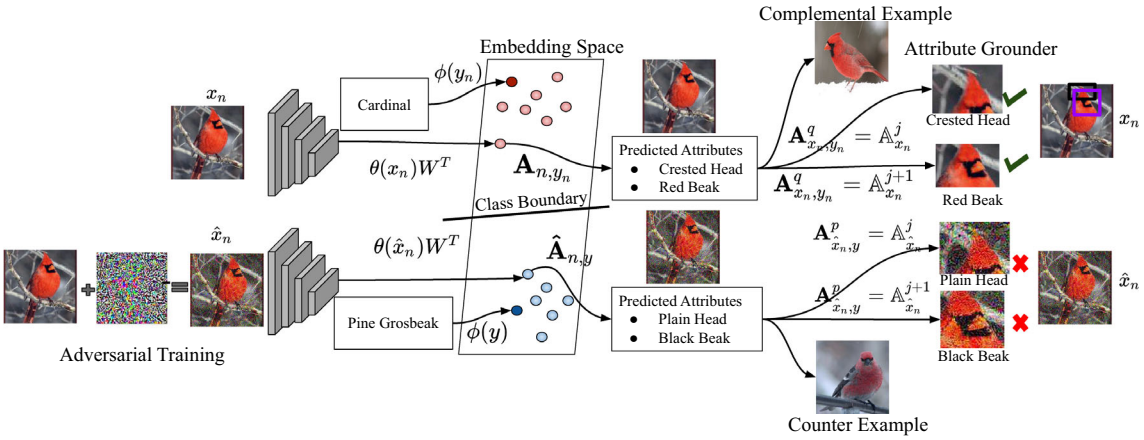

Fig. 3 Interpretable attribute prediction-grounding model. After an adversarial training step, image features of both clean $\theta(x_n)$ and adversarial images $\theta(\hat{x})$ are extracted using Resnet and mapped into attribute space $\phi(y)$ by learning the compatibility function $F(x_n, y_n; W)$ between image features and class attributes. Attributes predicted by attribute-based classifier $\mathbf{A}^q_{x_n,y_n}$ are grounded by matching them with attributes predicted by Faster-RCNN $\mathbb{A}^j_{x_n}$ for clean and adversarial images. Examples are selected based on attribute similarity between adversarial image and adversarial class images for visual explanations. Hence, clean image attributes lead to complemental explanations while, adversarial image attributes lead to counterfactual explanations

Pairwise ranking loss $\mathbb{L}(x_n, y_n, y)$ is used to learn the parameters (W):

$$\Delta(y_n, y) + \theta(x_n)^T W\phi(y_n) - \theta(x_n)^T W\phi(y) \tag{5}$$

At test time attributes are predicted for clean images by projecting image features on to the learned embedding space. It is given by:

$$\mathbf{A}_{n,y_n} = \theta(x_n)W \tag{6}$$

and for adversarial images by:

$$\hat{\mathbf{A}}_{n,y} = \theta(\hat{x}_n)W \tag{7}$$

The image is assigned the label of the nearest output class attributes $\phi(y_n)$.

3.4 Attribute grounding

Thereafter, we ground the predicted attributes on the images for better visual explanations using a pre-trained Faster-RCNN as in [4]. The pre-trained Faster-RCNN $\mathcal{F}(x_n)$ model predicts bounding boxes b^j. For each bounding box j in each image x_n it predicts a class $\mathbb{Y}^j_{x_n}$ and an attribute $\mathbb{A}^j_{x_n}$ [3].

$$b^j_{x_n}, \mathbb{A}^j_{x_n}, \mathbb{Y}^j_{x_n} = \mathcal{F}(x_n) \tag{8}$$

where j is the bounding box index.

Attribute selection for grounding As all the attributes predicted for an image cannot be visualized due to visual constraints. Therefore, we select the most discriminative attributes for grounding on the images. Attributes are selected based on the criterion that they change the most when the image is perturbed with the adversarial noise. For clean images we use:

$$q = \underset{i}{\text{argmax}}(\mathbf{A}^i_{n,y_n} - \phi(y^i)) \tag{9}$$

and for adversarial images we use:

$$p = \underset{i}{\text{argmax}}(\hat{\mathbf{A}}^i_{n,y} - \phi(y^i_n)). \tag{10}$$

where i is the attribute index, $\mathbf{A}^i_{n,y_n}$ and $\hat{\mathbf{A}}^i_{n,y}$ are attributes predicted by SJE for clean and adversarial images, respectively. $\phi(y^i)$, $\phi(y^i_n)$ indicate the counterclass and ground truth class attributes, respectively. q and p are indexes of the most discriminative attributes selected based on our criterion.

After selecting the most discriminative attributes predicted by SJE using Eqs. 9 and 10, we search for the selected attributes $\mathbf{A}^q_{x_n,y_n}$, $\mathbf{A}^p_{\hat{x}_n,y}$ in the attributes predicted by RCNN for each bounding box $\mathbb{A}^j_{x_n}$, $\mathbb{A}^j_{\hat{x}_n}$. When the attributes predicted by SJE and Faster-RCNN are matched, that is $\mathbf{A}^q_{x_n,y_n} = \mathbb{A}^j_{x_n}$, $\mathbf{A}^p_{\hat{x}_n,y} = \mathbb{A}^j_{\hat{x}_n}$ we ground them on their respective clean and adversarial images. As shown in Fig. 3, the attributes "Crested head" and "Red beak" are grounded on the image while "Plain head" and "Black beak" could not be grounded because there is no visual evidence present in the image for these attributes.

3.5 Example-based explanations

Besides providing attribute-based explanations we propose to provide counterexample-based explanations as shown in Fig. 3. We compare the results for example-based explanations by selecting examples randomly from the counterclass with examples selected based on attributes Fig. 13.

Example selection through attributes The procedure for example-based explanations using attributes is detailed in the Algorithm 1 and the results are shown in Figs. 12 and 13. Given clean images classified correctly and adversarial images misclassified and their predicted attributes, we search for attributes in the adversarial class which are most similar to the attributes of the adversarial image and select these images as counterexamples, i.e., a "Pine Grosbeak" image with the attributes "Plain head" and "Black beak" is selected as a counterexample Fig. 3.

Algorithm 1: Example Selection through Attributes

input : Adversarial images: $\hat{x}_{n,y}$, Clean images: x_{n,y_n}, Adversarial image attributes: $\hat{\boldsymbol{A}}_{n,y}$, Clean image attributes: $\mathbf{A}_{n,y_n}$, Adversarial classes: y

output: Selected examples from adversarial class: $x^s_{n,y}$

```
for each adversarial image x̂_{n,y} do
    Select all the images from adversarial class x_{n,y}
    for each image in adversarial class x_{n,y} do
        s = argmin_i || Â^i_{n,y} − A^i_{n,y} ||_2
    end
end
return x^s_{n,y}
```

3.6 Attribute analysis method

Finally, in this section we introduce our techniques for quantitative analysis on the predicted attributes.

Predicted attribute analysis: standard network In order to perform analysis on attributes in embedding space, we consider the images which are correctly classified without perturbations and misclassified with perturbations. Our aim is to analyze the change in attributes in embedding space to verify that attributes change with the change in the class.

We contrast the Euclidean distance between predicted attributes of clean and adversarial samples:

$$d_1 = d\{\mathbf{A}_{n,y_n}, \hat{\mathbf{A}}_{n,y}\} = \| \mathbf{A}_{n,y_n} - \hat{\mathbf{A}}_{n,y} \|_2 \tag{11}$$

with the Euclidean distance between the ground truth attribute vector of the correct and adversarial classes:

$$d_2 = d\{\phi(y_n), \phi(y)\} = \| \phi(y_n) - \phi(y)) \|_2 \tag{12}$$

where, $\mathbf{A}_{n,y_n}$ denotes the predicted attributes for the clean images classified correctly, and $\hat{\mathbf{A}}_{n,y}$ denotes the predicted attributes for the adversarial images misclassified with a standard network. The correct ground truth class attribute are referred to as $\phi(y_n)$ and adversarial class attributes are referred to as $\phi(y)$.

Predicted attribute analysis: robust network We compare the distances between predicted attributes of only adversarial images that are classified correctly with the help of an adversarially robust network $\hat{\mathbf{A}}^r_{n,y_n}$ and classified incorrectly with a standard network $\hat{\mathbf{A}}_{n,y}$:

$$d_1 = d\{\hat{\mathbf{A}}^r_{n,y_n}, \hat{\mathbf{A}}_{n,y}\} = \| \hat{\mathbf{A}}^r_{n,y_n} - \hat{\mathbf{A}}_{n,y} \|_2 \tag{13}$$

with the distances between the ground truth class attributes $\phi(y_n)$ and ground truth adversarial class attributes $\phi(y)$:

$$d_2 = d\{\phi(y_n), \phi(y)\} = \| \phi(y_n) - \phi(y)) \|_2 \tag{14}$$

3.7 Implementation details

Image features and adversarial examples We extract image features and generate adversarial images using the fine-tuned Resnet-152. Adversarial attacks are performed using the basic iterative method with epsilon ϵ values 0.01, 0.06 and 0.12. The l_∞ norm is used as a similarity measure between clean input and the generated adversarial example. In order to generate adversarial examples for untargeted attacks the algorithm perturbs the images such that they get misclassified into any alternative counter class. In order to generate adversarial examples for targeted attacks we direct adversarial examples to be misclassified into randomly selected classes.

Adversarial training As for adversarial training, we repeatedly computed the adversarial examples while training the fine-tuned Resnet-152 to minimize the loss on these examples. We generated adversarial examples using the projected gradient descent method. This is a multi-step variant of FGSM with epsilon ϵ values 0.01, 0.06 and 0.12, respectively, for adversarial training as in [27].

Attribute prediction and grounding At test time the image features are projected onto the learned attribute space and attributes per image are predicted. The image is assigned with the label of the nearest ground truth attribute vector. Since we do not have ground truth part bounding boxes for any of the attribute datasets, the predicted attributes are grounded by using Faster-RCNN pre-trained on the Visual Genome Dataset [22] .

4 Experiments and results

4.1 Datasets

We experiment on three datasets, Animals with Attributes 2 (AwA) [24], Large attribute (LAD) [41] and Caltech UCSD Birds (CUB) [31]. AwA contains 37322 images (22206 training/5599 validation/9517 test) with 50 classes and 85 attributes per class. LAD has 78017 images (40957 training/13653 validation/23407 test) with 230 classes and 359 attributes per class. CUB consists of 11,788 images (5395 training/599 validation/5794 test) assigned to 200 fine-grained categories of birds with 312 attributes per class. All three datasets contain real-valued class attributes representing the degree of presence of an attribute in a class. For the qualitative analysis with grounding we select 50 attributes that change their value most for the CUB, 50 attributes for AWA, and 100 attributes for the LAD dataset. They are selected by Eqs. 9 and 10, since it is difficult for humans to understand all the attributes grounded on the images.

The Visual Genome Dataset [22] is used to train the Faster-RCNN model which extracts the bounding boxes using 1600 object and 400 attribute annotations. Each bounding box is associated with an attribute and the class, e.g. a brown bird.

4.2 Comparing general and attribute-based classifiers

In the first experiment, we compare the general classifier $f(x_n)$ and the attribute-based classifier $\mho(x_n)$ in terms of the classification accuracy on clean images to see whether the attribute-based classifier performs equally well.

We find that, the attribute-based and general classifier accuracies are comparable for AWA (general: 93.53, attribute-based: 93.83). The attribute-based classifier accuracy is slightly higher for LAD (general: 80.00, attribute-based: 82.77), and lower for CUB (general: 81.19, attribute-based: 76.90) dataset. The overall impression is that both general and attribute-based classifiers perform equally well.

4.3 Attribute-based explanations: standard network

In the second experiment we study the change in attributes with a standard network to demonstrate that by introducing perturbations in the images the attribute values change such that the class of the image changes to the counterclass and hence provide intuitive counterexplanations. We study the change in attribute values both when the counterclass is any other class, i.e., untargeted, as well as when we direct the image into a specific class, i.e., targeted.

4.3.1 By performing classification based on attributes

Untargeted attacks With untargeted adversarial attacks, the accuracy of both the general and attribute-based classifiers drops with the increase in perturbations see Fig. 4 (blue curves). The drop in accuracy of the general classifier for the fine-grained CUB-dataset is higher as compared to the coarse-grained AWA dataset. For example, at $\epsilon = 0.01$ for the CUB dataset the general classifier's accuracy drops from 81% to 31% , while for the AWA dataset it drops from 93% to 70% and for LAD dataset it drops from 80% to 50%. However, compared to the general classifier the drop in accuracy with the attribute-based classifier for CUB dataset is less $\approx$ 20%. For the coarse-grained datasets AWA and LAD the drop is almost the same for both attribute-based and general classifiers. The limited drop in accuracy for the CUB dataset with the attribute-based classifier when compared to the general classifier, is attributed to the fact that for fine-grained datasets there are many attributes common among classes. Therefore, in order to misclassify an image a significant number of attributes need to change their values. For a coarse-grained dataset, changing a few attributes is sufficient for misclassification. Overall, the drop in the accuracy due to the perturbation demonstrates that the attribute values change toward those that belong to the new class. Hence, attributes explain the misclassifications into the counterclasses well. This also concludes that attributes contain the crucial characteristics for discrimination between classes.

Targeted attacks In the untargeted attacks the algorithm misclassifies the image into any alternative class which could be a closer class, i.e., a class with the majority of attribute values same as the ones from the ground truth class. In contrast, targeted adversarial attacks force the image to be misclassified into a randomly selected desired class which could be far away from the ground truth class, i.e., the attribute values between both classes differ significantly, hence making the targeting into this class difficult. We evaluate our method for misclassification into a closer class as well as for a distant class.

The accuracy of both general as well as attribute-based classifiers drop with the increase in perturbations see Fig. 5 (blue curves). However, compared to the drop in performance with untargeted attacks the drop with targeted attacks is lower see Figs. 4 and 5 (blue curves). This is due to the fact that in untargeted attacks the images are misclassified into closer classes while with the targeted attacks images get misclassified into distant classes.

By contrasting the drop in the accuracy of the general classifier between three datasets using targeted attacks we observe that the fine grained CUB-dataset leads to a higher drop in the performance as compared to the AWA, and LAD datasets Fig. 5 (blue solid curves). Although the drop with

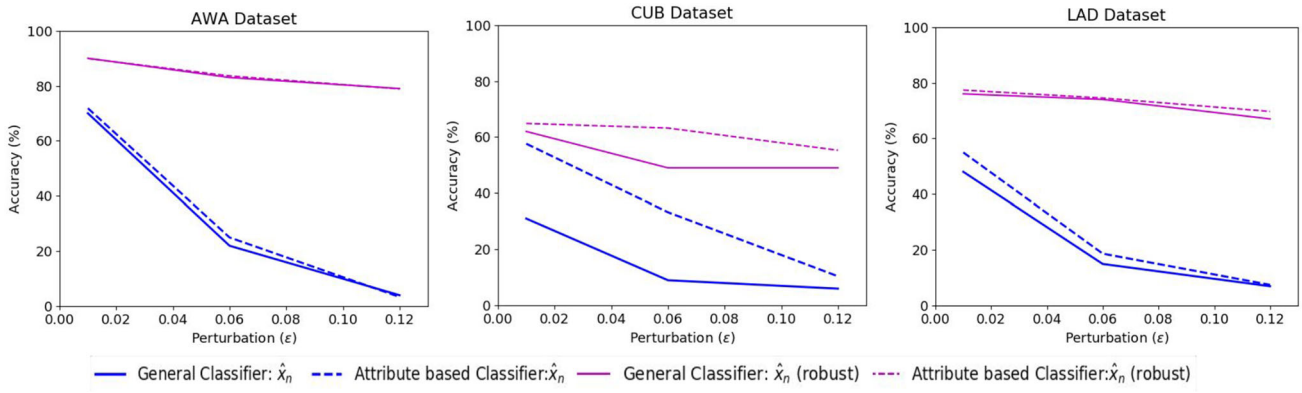

Fig. 4 Untargeted Attacks: Comparing the accuracy of the general classifier and the attribute-based classifier for adversarial examples generated with untargeted attacks to investigate the change in attributes. We evaluate both classifiers by extracting features from a standard network and the adversarially robust network. The drop in the performance with the increase in the level of perturbations shows that the attributes start pointing toward the counter classes (blue curves). The improvement in the performance with robustification shows that with an adversarially robustified network the attributes again start pointing toward the ground truth class (purple curves) (color figure online)

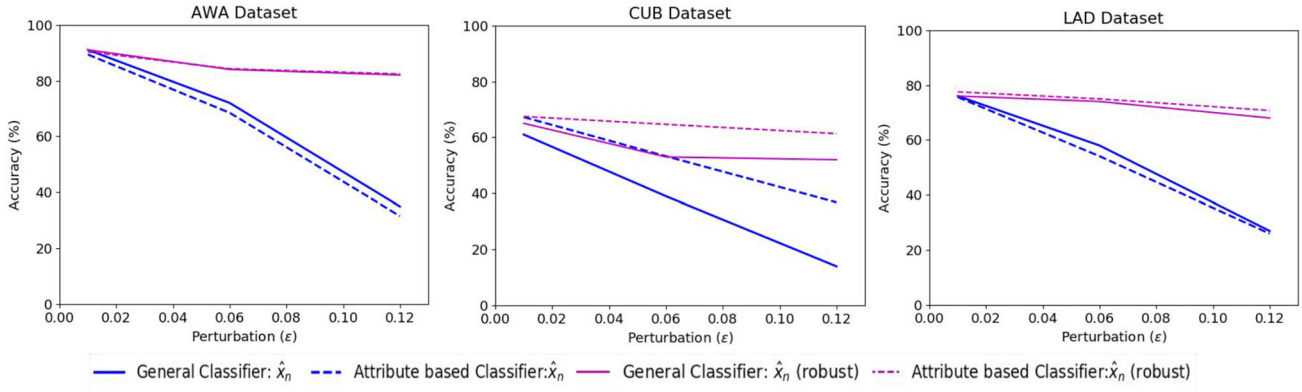

Fig. 5 Targeted Attacks: Comparing the accuracy of the general classifier and the attribute-based classifier for adversarial examples generated with targeted attacks to investigate the change in attributes. We evaluate both classifiers by extracting features from a standard network and the adversarially robust network. The drop in the performance with the increase in the level of perturbations shows that the attributes start pointing toward the counter classes (blue curves). However, the drop is not significant when compared to untargeted attacks. Similarly, with the adversarial robustness the performance improves and the attributes start pointing toward the ground truth class, however the improvement is also not as significant as for the untargeted attacks (purple curves) (color figure online)

targeted attacks is lower than untargeted attacks but the overall behavior in the drop is the same for both untargeted and targeted attacks. For instance, at $\epsilon = 0.06$ the accuracy drops from 81% to 39% for CUB-dataset, while for AWA dataset it drops from 93% to 72% and for LAD dataset it drops from 80% to 58%. While the drop in the accuracy with the attribute-based classifier for CUB-dataset reduced to almost half, i.e., $\approx$ 23% and increased for AWA and LAD dataset, i.e., $\approx$ 25% and $\approx$ 29%, respectively. Similar to the general classifier attribute based classifier for targeted attacks also shows the same behavior as attribute based-classifier for untargeted attacks. Hence, this further supports our argument that for fine grained datasets as there are numerous attributes common among the classes therefore we need to change many of them in order to change the class and provide explanations based on the attributes. While, for the coarse grained datasets only by changing a few attributes we can cause misclassification and explain it.

Overall, the lack in the drop of performance for an attribute based classifier with the targeted attacks as compared to untargeted attacks shows that the change in attribute values towards the counterclass is less significant with the targeted attacks. Hence, attribute values with untargeted attacks provide better counterexplanations than attribute values with the targeted attacks.

4.3.2 By computing distances in the embedding space

We contrast the Euclidean distance between predicted attributes of clean and adversarial samples using Eqs. 11 and 12. The

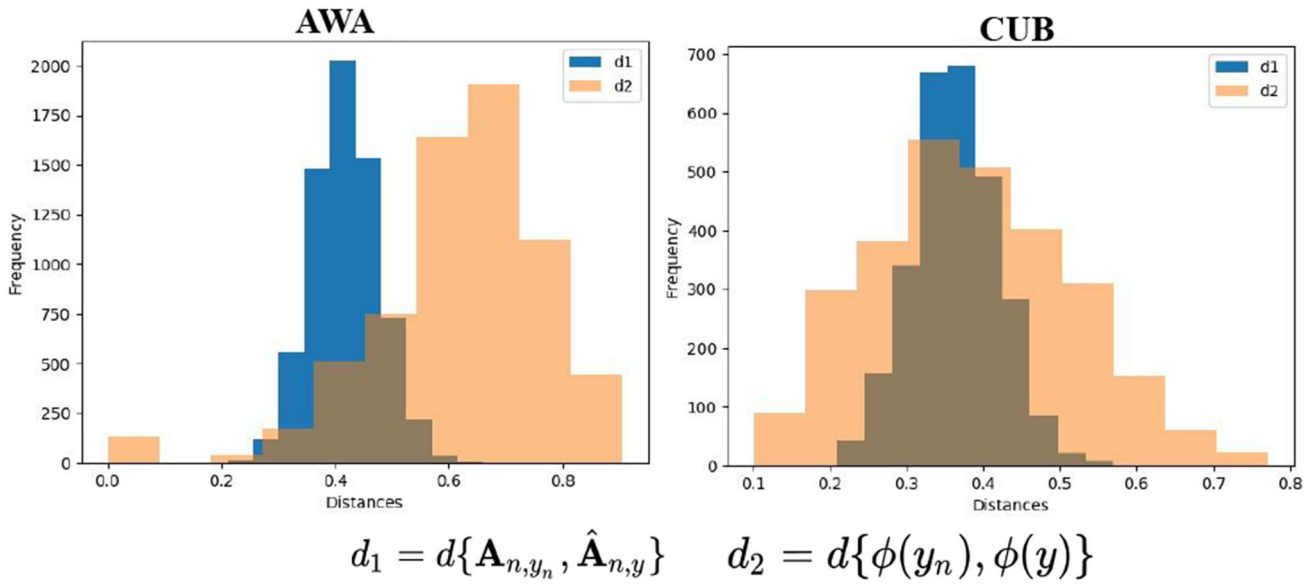

Fig. 6 Attribute value distance plots for clean and adversarial images with a standard network. The complete overlap for the CUB-dataset shows that fine-grained datasets require change in significant no of attribute to change the class. While the small overlap for the coarse-grained AWA dataset shows that the change in a few attributes is sufficient to change the class

results are shown in Fig. 6. We observe that for the AWA dataset the distances between the predicted attributes for adversarial and clean images d_1 are smaller than the distances between the ground truth attributes of the respective classes d_2. The closeness in predicted attributes for clean and adversarial images as compared to their ground truths shows that attribute values change towards the wrong class but not completely. This is due to the fact that for coarse classes only a small change in attribute values is sufficient to change the class.

The fine-grained CUB-dataset behaves differently. The overlap between d_1 and d_2 distributions demonstrates that attributes of images belonging to fine-grained classes change significantly as compared to images from coarse categories. As the fine-grained classes are closer to one another and many attributes are common among fine-grained classes. Thus it requires to change the attributes significantly to cause misclassification. Hence, for the coarse-grained dataset, the attributes change minimally, while for the fine-grained dataset they change significantly.

4.3.3 Qualitative analysis

Untargeted attacks We observe in Fig. 8 that the most discriminative attributes for the clean images are coherent with the ground truth class however, for adversarial images they are coherent with the wrong class thus explaining the wrong class. For example "red head, black wing, black eye" attributes are responsible for the classification of clean image into correct class and when the value of "red head" attribute

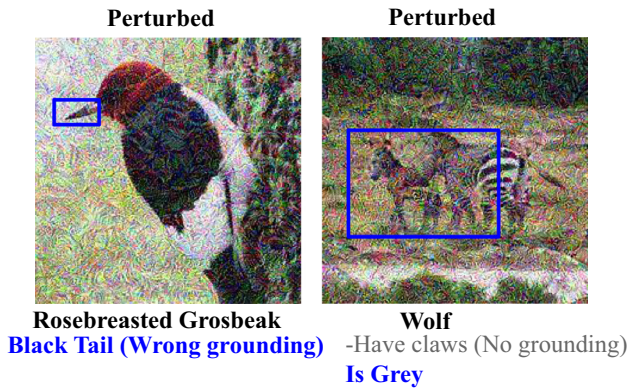

Fig. 7 Explanation of a wrong classification due to wrong or missing attribute grounding. For perturbed images attributes either get grounded on wrong spots or are missing because their visual evidence is absent in the image. (Perturbations magnified)

decreases and "grey beak, white underparts" increases the image gets misclassified into wrong class. Figure 7 reveals the results for the groundings on perturbed images. The attributes which are not related to the correct class, the ones that are related to the counterclass cannot get grounded or get grounded at the wrong spots in the image as there is no visual evidence that supports the presence of these attributes. For example "black tail" is related to the counterclass and is not present in the adversarial image. Hence, black tail" got wrongly grounded. This indicates that attributes for the clean images correspond to the ground truth class and for adversarial images correspond to the counterclass. Addition-

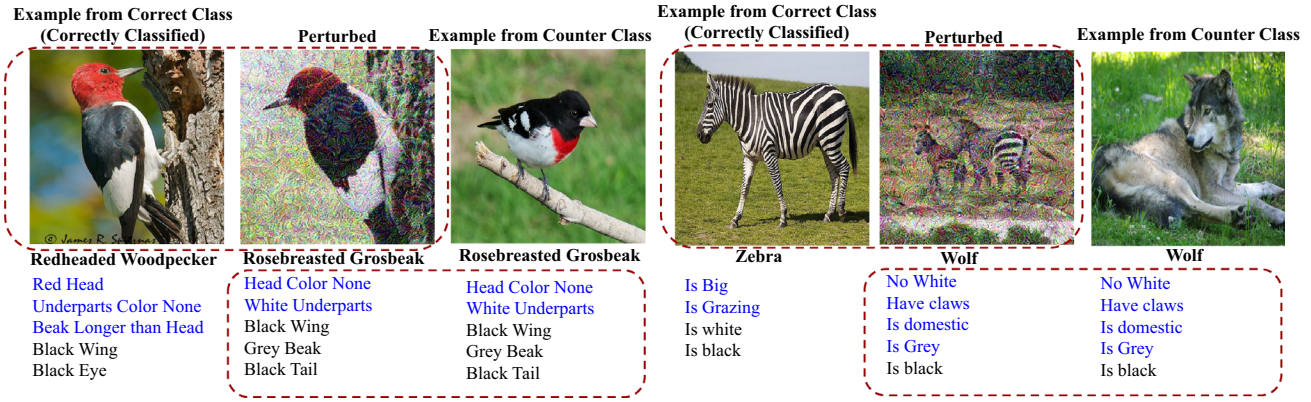

Fig. 8 Untargeted: Qualitative analysis for change in attributes due to directed perturbations with a standard network. The attributes are ranked by importance for classification. Most discriminative attributes for clean images correspond to the ground truth class while, those for the perturbed image they compatible with the counter class thus explaining the misclassification. (Perturbations magnified for better visibility)

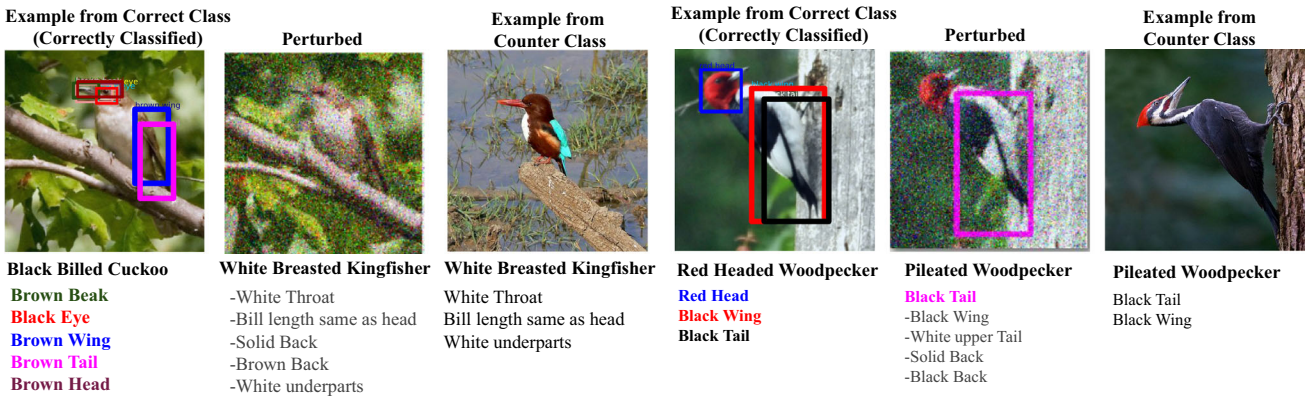

Fig. 9 Targeted: Qualitative analysis for change in attributes due to directed perturbations with a standard network. The attributes are ranked by importance for classification. Grounded attributes are color coded for the visibility. Those in gray color could not be grounded. Those attributes common among ground truth and counter class are grounded while those for which no visual evidence is found in the image could not be grounded on the perturbed image hence, indicating the change in the class. (Perturbations magnified for better visibility.) (color figure online)

ally, only those attributes common among both the counterand the ground truth classes get grounded on adversarial images.

Hence, our method provides explanations for both fine and coarse-grained classifications when the images get misclassified into similar classes or dissimilar classes.

Targeted attacks Figure 9 reveals the results for grounding the attributes when the images are misclassified with targeted attacks. As in the targeted attacks we direct images into random classes we observe that images get misclassified into visually dissimilar classes. The attributes predicted for perturbed images also correspond to visually dissimilar counterclasses. Hence, it becomes difficult to ground predicted attributes on perturbed images because there is no visual evidence present for those attributes in the images. For instance in fig. 9 first example, "White Throat", "Bill length same as head" and "Solid Back" were responsible for misclassification into the "White breasted kingfisher" class, but as there is no visual evidence available for these attributes in the image originally belonging to "Black billed Cuckoo" class therefore, none of the attributes could be grounded on the perturbed image. Hence, our results show that the visual explanations provided by untargeted perturbations are much more useful for human understanding as compared to targeted perturbations.

4.4 Attribute-based explanations: robust network

We perform the same experiments with a robust network to study the change in attribute values such that the class of the perturbed image changes back to the ground truth class.

4.4.1 By performing classification based on attributes

Untargeted attacks Our evaluation on the standard and adversarially robust networks shows that the classification accuracy improves for the adversarial images when adversarial training is used to robustify the network Fig. 4 (purple curves). For example, in Fig. 4 for AWA the accuracy of the general classifier improved from 70% to 92% and for LAD it improved from 50% to 78% for adversarial attack with $\epsilon = 0.01$. As expected for the fine-grained CUB-dataset the improvement is $\approx 31\%$ higher than the AWA and LAD datasets. However, for the attribute-based classifier, the improvement in accuracy for AWA ($\approx 18\%$) is almost double and for LAD ($\approx 22\%$) almost triple that of the CUB-dataset ($\approx 7\%$). This demonstrates that, attributes retain their discriminative power for explanations with the standard as well as robust networks.

Targeted attacks Results for the performance of standard and adversarially robust networks against targeted attacks show that the performance of the network improves for adversarial images when tested on an adversarially robust network Fig. 5 (purple curves). Different from untargeted attacks for targeted attacks, the improvement in the performance is not significant. For example, in Fig. 5 at $\epsilon = 0.06$ for AWA dataset the accuracy improved to $\approx 12\%$, for CUB it improved to $\approx 14\%$ and for LAD dataset it improved to $\approx 16\%$ while with untargeted attacks the improvement in the accuracy at $\epsilon = 0.06$ is more than double of that with targeted attacks. This shows that when images are misclassified into visually dissimilar classes it becomes difficult to correctly classify them with robustification as compared to images misclassified into visually similar classes.

Similarly, for attribute-based classifier the improvement in the accuracy is less for targeted attacks as compared to the untargeted attacks Fig. 5 (purple dotted curves). The overall behavior in the improvement of performance for each dataset with targeted attacks is similar to that of untargeted attacks. For instance, at $\epsilon = 0.06$ the improvement in the accuracy for the CUB-dataset is the least $\approx 11\%$ following AWA $\approx 16\%$ and LAD $\approx 21\%$ datasets. This supports our argument that in order to change the class of fine grained images more number of attributes need to be changed. Overall, our results reveal that even for an adversarially robustified network untargeted attacks provide better explanations as compared to targeted attacks.

4.4.2 By computing distances in the embedding space

We also compare the euclidean distance between predicted attributes for only adversarial images in the presence of a standard network and a robust network as shown in Fig. 10. The results reveal that with only adversarial images on robust and standard networks we observe the same distance distribution as in Fig. 6. Thus, attributes explain the correct classification of adversarial images in the presence of the robust network.

4.4.3 Qualitative analysis

Finally, our analysis with correctly classified images by the adversarially robust network shows that, adversarial images and their predicted attributes with the robust network behave like clean images and their predicted attributes as shown in Fig. 11. This also demonstrates that the attributes for adversarial images classified correctly with the robust network still retain their discriminative power and provide complementary explanations.

4.5 Example-based explanations

In the final experiment we demonstrate our visual example and counterexample-based explanations when the attribute values change with directed perturbations. For instance in Fig. 12 when an image is classified correctly besides explaining the classification decision with attributes we enhance our explanations with the complemental example retrieved based on these attributes. Similarly, when an image is misclassified into a counter class we also enhance our attribute-based explanations by retrieving an image from the counter class.

Figure 13 reveals the importance of counterexample selection through attributes. In this example both the clean images in first and second row belong to the same class the "Mallard". However, the clean image in the first row is male Mallard and in the second row is female Mallard, they differ visually. Similarly, the male and female birds of the counterclass "Redbreasted Merganser" also differ visually. The results for the examples retrieval for both male and female mallard show that, when images are retrieved through attributes for the male Mallard the retrieved images are male Redbreasted Merganser, while for the female Mallard the retrieved images through attributes are female Redbreasted Merganser. However, when we retrieve the images randomly from the counterclass then the visual similarity can not be ensured. Hence, our attribute-based example selection method selects the visually similar examples to provide the distinction between a clean image and a counter image from the counter class under the presence of intra-class variation.

5 Discussion and conclusion

In this work we focused on providing the understanding of neural networks decisions by exploiting counterattributes as well as counterexamples which lead to the misclassification in the counterclass.

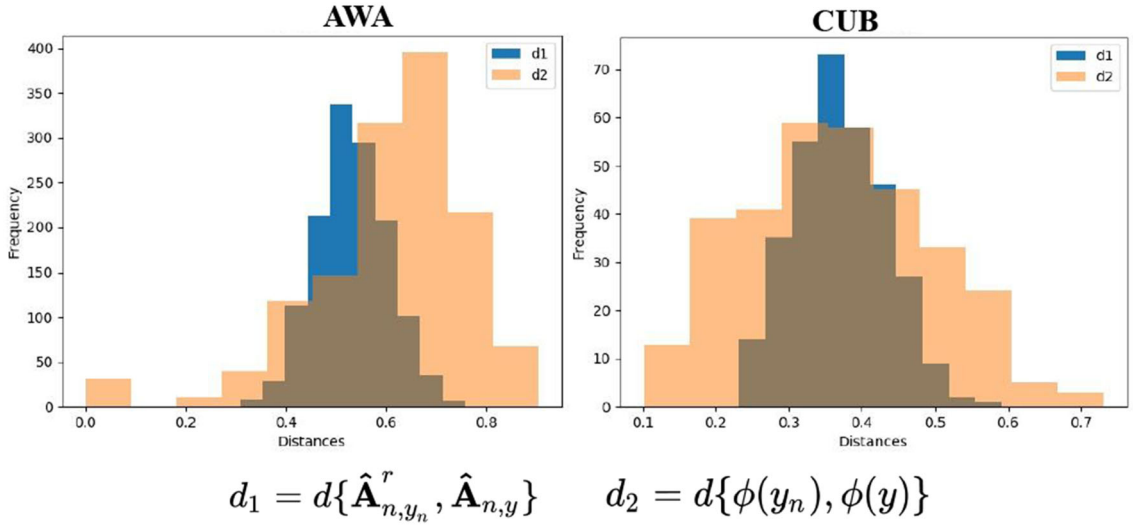

Fig. 10 Attribute value distance plots for only adversarial images with and without a robust network. The similarity with the plots in Fig. 6 shows that adversarial image attributes in the presence of a robust network indicate to the ground truth class

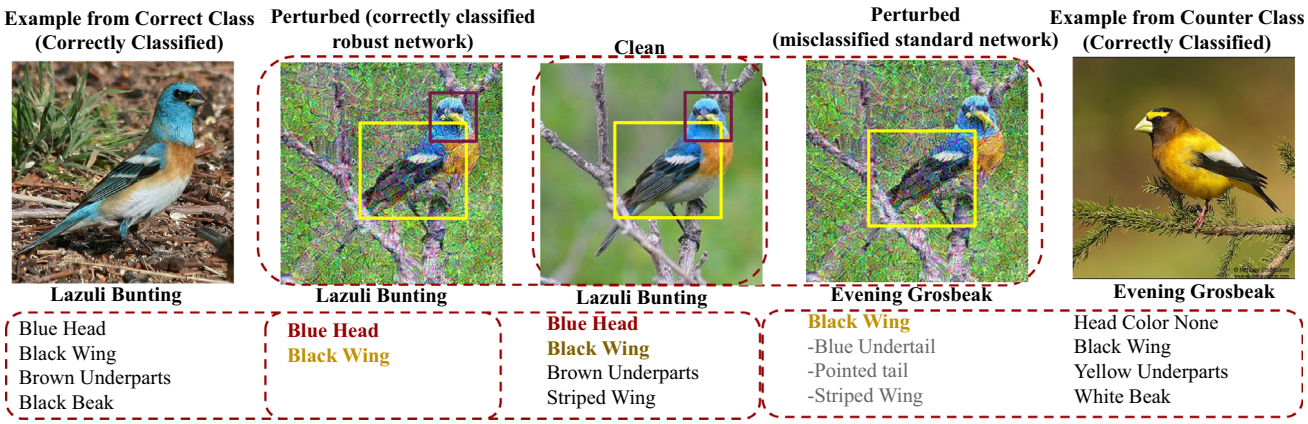

Fig. 11 Qualitative analysis for change in attributes due to directed perturbations with a robust network. The attributes are ranked by importance for the classification decision, the grounded attributes are color coded for visibility (the ones in gray could not be grounded). The overlap between the attributes of adversarial image with a robust network and a clean image with a standard network shows that with a robust network attributes change back to the ground truth class. (Perturbations magnified for better visibility)

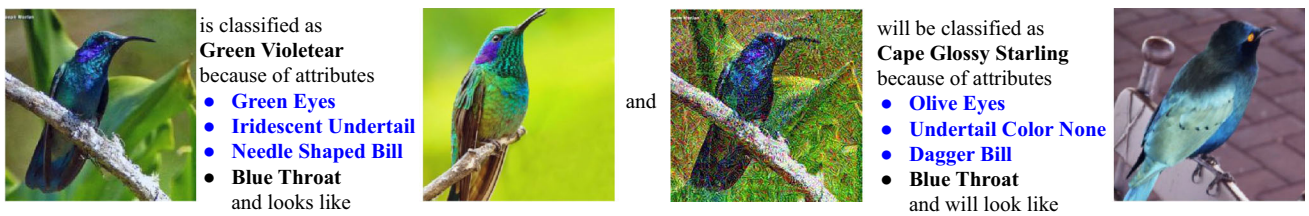

Fig. 12 Qualitative analysis for Example-based explanations. Note that when "green eyes, needle shaped bill" changes to "olive eyes, dagger bill" the class of the image changes. These attributes are also complemented with the image-based examples retrieved with these attributes. (Perturbations magnified for better visibility)

Firstly, we showed that attribute-based classifiers perform equally well as direct classifiers. We also showed that the importance of attributes for providing explanations is higher for the fine-grained classification as compared to coarse-grained classification because the distinction between two coarse-grained classes can be made through a single attribute

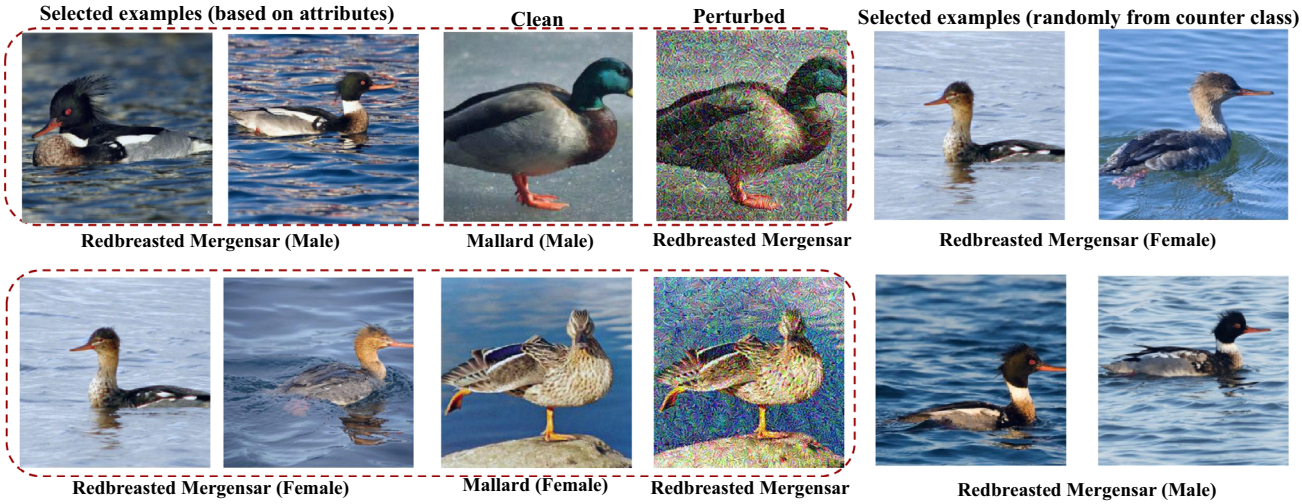

Fig. 13 Qualitative analysis for Example-based explanations. Note that both "Mallard" and "Redbreasted Mergensar" classes have intra-class variability as the male and female birds in both classes look visually different. When we use attributes for retrieving image examples, male Mallard retrieves male Redbrested Mergensar and female Mallard retrieves female Redbrested Merganser thus incorporating the intra-class variability. (Perturbations magnified for better visibility)

as compared to the fine-grained classes which require numerous attributes for distinction between them.

Secondly, we demonstrated that by introducing adversarial perturbations in the images we were able to change the attribute values to those of counterclass attributes and hence provided counterattribute-based explanations. Our results showed that these attributes contain crucial characteristics for the discrimination between classes.

Thirdly, we repeated all the experiments for the images with perturbations introduced through targeted attacks. Our results showed that, our attribute-based explanations work better with untargeted attacks as compared to the targeted attacks.

We also showed that when a network is robustified against adversarial perturbations the predicted attribute values for the perturbed images start indicating back towards the correct class which further confirmed our attribute-based explanations.

Finally, we demonstrated our attribute-based explanations by providing causal reasoning "because the image contains these attributes therefore it is classified into this class". We also assisted our counterattribute-based explanations with counterexamples selected based on predicted attributes and showed that our method selected most precise and illustrative examples even in the presence of intra-class variations.

Hence, we conclude that attributes provide intuitive factual and in the presence of perturbations counterfactual human understandable explanations especially for fine grained classification. These explanations could also be enhanced by retrieving visual examples through them. Attributes retain their best discriminative power in the presence of untargeted attacks with standard as well as robustified networks.

References

1. Abbasnejad E, Teney D, Parvaneh A, Shi J, Hengel Avd (2020) Counterfactual vision and language learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 10044-10054
2. Akata Z, Reed S, Walter D, Lee H, Schiele B (2015) Evaluation of output embeddings for fine-grained image classification. In: CVPR
3. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR
4. Anne Hendricks L, Hu R, Darrell T, Akata Z (2018) Grounding visual explanations. In: ECCV
5. Browne K, Swift B (2020) Semantics and explanation: why counterfactual explanations produce adversarial examples in deep neural networks. arXiv preprint, arXiv:2012.10076

6. Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE Computer Society, pp 39–57
7. Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: SP. IEEE
8. Carlini N, Wagner D (2018) Audio adversarial examples: targeted attacks on speech-to-text. In: 2018 IEEE security and privacy workshops (SPW). San Francisco, CA, USA, pp 1–7. https://doi.org/10.1109/SPW.2018.00009
9. Dong Y, Su H, Zhu J, Zhang B (2017) Improving interpretability of deep neural networks with semantic information. In: CVPR
10. Du M, Liu N, Hu X (2019) Techniques for interpretable machine learning. Commun. ACM 63(1):68–77. https://doi.org/10.1145/3359786
11. Edwards L, Veale M (2017) Slave to the algorithm: why a right to an explanation is probably not the remedy you are looking for. Duke L Tech Rev 16:18
12. Fong RC, Vedaldi A (2017) Interpretable explanations of black boxes by meaningful perturbation. In: 2017 IEEE International conference on computer vision (ICCV), Venice, Italy, pp 3449–3457. https://doi.org/10.1109/ICCV.2017.371
13. Goyal Y, Wu Z, Ernst J, Batra D, Parikh D, Lee S (2019) Counterfactual visual explanations. In: International conference on machine learning, PMLR. pp 2376–2384
14. Gulshad S, Smeulders A (2020) Explaining with counter visual attributes and examples. In: Proceedings of the 2020 international conference on multimedia retrieval. pp 35–43
15. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z (2019) XAI-explainable artificial intelligence. Sci Robot 4(37):eaay7120. https://doi.org/10.1126/scirobotics.aay7120
16. Hendricks LA, Akata Z, Rohrbach M, Donahue J, Schiele B, Darrell T (2016) Generating visual explanations. In: ECCV. Springer
17. Hendricks LA, Hu R, Darrell T, Akata Z (2018) Generating counterfactual explanations with natural language. In: ICML workshop on human interpretability in machine learning. pp 95–98
18. Hsieh CY, Yeh CK, Liu X, Ravikumar P, Kim S, Kumar S, Hsieh C.J (2020) Evaluations and methods for explanation through robustness analysis. https://openreview.net/forum?id=Hye4KeSYDr
19. Ignatiev A, Narodytska N, Marques-Silva J (2019) On relating explanations and adversarial examples. In: Advances in neural information processing systems. Association for Information Systems, pp 15857–15867
20. Jiang L, Liu S, Chen C (2019) Recent research advances on interactive machine learning. J Vis 22:401–417. https://doi.org/10.1007/s12650-018-0531-1
21. Kanehira A, Harada T (2019) Learning to explain with complemental examples. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 8603–8611
22. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. IJCV
23. Kurakin A, Goodfellow I, Bengio S (2017) Adversarial examples in the physical world. ICLR workshop
24. Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: CVPR. IEEE
25. Liu S, Kailkhura B, Loveland D, Han Y (2019) Generative counterfactual introspection for explainable deep learning. arXiv preprint arXiv:1907.03077
26. Loyola-González O (2019) Black-box vs. white-box: understanding their advantages and weaknesses from a practical point of view. IEEE Access 7:154096–154113. https://doi.org/10.1109/ACCESS.2019.2949286
27. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2018) Towards deep learning models resistant to adversarial attacks. ICLR
28. Moosavi-Dezfooli SM, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: CVPR
29. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A (2016) The limitations of deep learning in adversarial settings. In: EuroS&P. IEEE
30. Park DH, Hendricks LA, Akata Z, Schiele B, Darrell T, Rohrbach M (2018) Multimodal explanations: justifying decisions and pointing to the evidence. In: CVPR
31. Reed S, Akata Z, Lee H, Schiele B (2016) Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 49–58
32. Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you?: Explaining the predictions of any classifier. In: ACM SIGKDD
33. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: ICCV
34. Shrikumar A, Greenside P, Kundaje A (2017) Learning important features through propagating activation differences. In: ICML
35. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint, arXiv:1312.6034
36. Su J, Vargas DV, Sakurai K (2019) One pixel attack for fooling deep neural networks. TEVC
37. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. ICLR
38. Tsipras D, Santurkar S, Engstrom L, Turner A, Madry A (2019) Robustness may be at odds with accuracy. In: International conference on learning representations
39. Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: automated decisions and the gdpr. Harv JL Tech 31:841
40. Zhang T, Zhu Z (2019) Interpreting adversarially trained convolutional neural networks. In: International conference on machine learning. PMLR, pp 7502–7511
41. Zhao B, Fu Y, Liang R, Wu J, Wang Y, Wang Y (2019) A large-scale attribute dataset for zero-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops
42. Zintgraf LM, Cohen TS, Adel T, Welling M (2017) Visualizing deep neural network decisions: prediction difference analysis. ICLR