EXPLANATION BY PROGRESSIVE EXAGGERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As machine learning methods see greater adoption and implementation in high stakes applications such as medical image diagnosis, the need for model interpretability and explanation has become more critical. Classical approaches that assess feature importance (*e.g.*, saliency maps) do not explain *how* and *why* a particular region of an image is relevant to the prediction. We propose a method that explains the outcome of a classification black-box by gradually *exaggerating* the semantic effect of a given class. Given a query input to a classifier, our method produces a progressive set of plausible variations of that query, which gradually change the posterior probability from its original class to its negation. These counter-factually generated samples preserve features unrelated to the classification decision, such that a user can employ our method as a "tuning knob" to traverse a data manifold while crossing the decision boundary. Our method is model agnostic and only requires the output value and gradient of the predictor with respect to its input.

1 INTRODUCTION

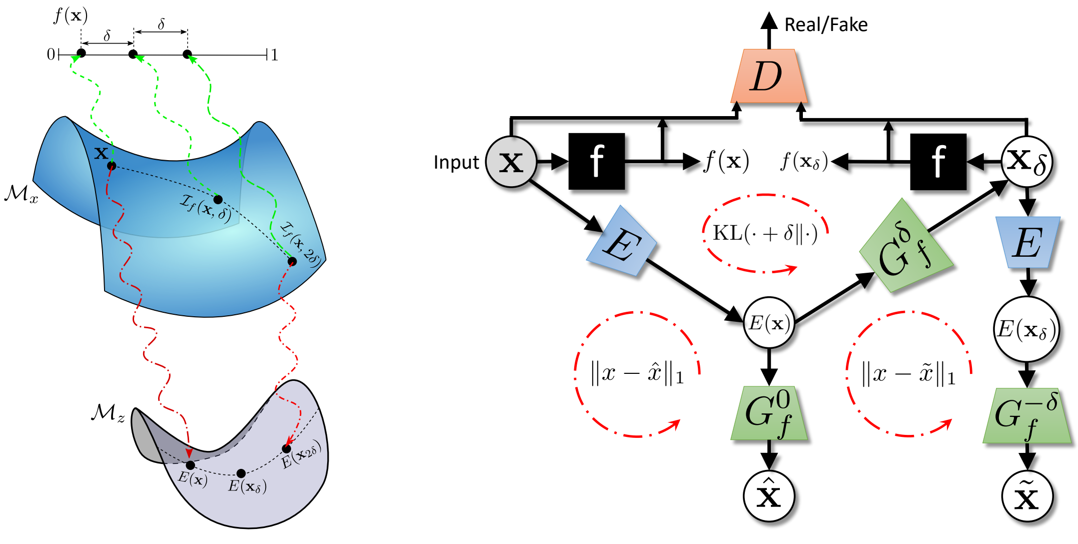
With the explosive adoption of deep learning for real-world applications, explanation and model interpretability have received substantial attention from the research community (Kim, 2015; Doshi-Velez & Kim, 2017; Molnar, 2018; Guidotti et al., 2019). Explaining an outcome of a model in high stake applications, such as medical diagnosis from radiology images, is of paramount importance to detect hidden biases in data (Cramer et al., 2018), evaluate the fairness of the model (Doshi-Velez & Kim, 2017), and build trust in the system (Glass et al., 2008). For example, consider evaluating a computer-aided diagnosis of Alzheimer's disease from medical images. The physician should be able to assess whether or not the model pays attention to age-related or disease-related variations in an image in order to trust the system. Given a query, our model provides an explanation that gradually *exaggerates* the semantic effect of one class, which is equivalent to traversing the decision boundary from side to another.

Although not always clear, there are subtle differences between *interpretability* and *explanation* (Turner, 2016). While the former mainly focuses on building or approximating models that are locally or globally interpretable (Ribeiro et al., 2016), the latter aims at explaining a predictor a-posteriori. The explanation approach does not compromise the prediction performance. However, a rigorous definition for what is a good explanation is elusive. Some researchers focused on providing feature importance (*e.g.*, in the form of a heatmap (Selvaraju et al., 2017)) that influence the outcome of the predictor. In some applications (*e.g.*, diagnosis with medical images) the causal changes are spread out across a large number of features (*i.e.*, large portions of the image are impacted by a disease). Therefore, a heatmap may not be informative or useful, as almost all image features are highlighted. Furthermore, those methods do not explain *why* a predictor returns an outcome. Others have introduced local occlusion or perturbations to the input (Zhou et al., 2014; Fong & Vedaldi, 2017) by assessing which manipulations have the largest impact on the predictors. There is also recent interest in generating counterfactual inputs that would change the black box classification decision with respect to the query inputs (Goyal et al., 2019; Liu et al., 2019). Local perturbations of a query are not guaranteed to generate realistic or plausible inputs, which diminishes the usefulness of the explanation, especially for end users (*e.g.*, physicians). We argue that the explanation should depend not only on the predictor function but also on the data. Therefore, it is reasonable to train a model that learns from data as well as the black-box classifier (*e.g.*, (Chang et al., 2018; Dabkowski & Gal, 2017; Fong & Vedaldi, 2017)).

Our proposed method falls into the local explanation paradigm. Our approach is model agnostic and only requires access to the predictor values and its gradient with respect to the input. Given a query input to a black-box, we aim at explaining the outcome by providing *plausible* and *progressive* variations to the query that can result in a change to the output. The plausibility property ensures that perturbation is natural-looking. A user can employ our method as a ”tuning knob” to progressively transform inputs, traverse the decision boundary from one side to the other, and gain understanding about how the predictor makes a decision. We introduce three principles for an explanation function that can be used beyond our application of interested. We evaluate our method on a set of benchmarks as well as real medical imaging data. Our experiments show that the counterfactually generated samples are realistic-looking and in the real medical application, satisfy the external evaluation. We also show that the method can be used to detect bias in training of the predictor.

2 Method

Consider a *black box* classifier that maps an input space $\mathcal{X}$ (*e.g.*, images) to an output space $\mathcal{Y}$ (*e.g.*, labels). In this paper, we consider binary classification problems where $\mathcal{Y} = \{-1, +1\}$. To model the black-box, we use $f(\mathbf{x}) = \mathbb{P}(y|\mathbf{x})$ to denote the posterior probability of the classification. We assume that f is a differentiable function and we have access to its value as well as its gradient with respect to the input $\nabla_{\mathbf{x}} f(\mathbf{x})$.



(a)

(b)

Figure 1: **(a)** The schematic of the method: f is the black-box function producing the posterior probability. $\mathcal{I}_f(\mathbf{x}, \delta)$ is an explainer function for f, which shifts the value of $f(\mathbf{x})$ by δ. The $E(\cdot)$ is an encoder that maps the data manifold $\mathcal{M}_x$ to the embedding manifold $\mathcal{M}_z$. $\mathbf{x}_\delta$ is an abbreviation for $\mathcal{I}_f(\mathbf{x}, \delta)$. **(b)** The architecture of our model: E is the encoder, G_f^δ denotes the conditional generator $G(\cdot, c_f(x, \delta))$, f is the black-box and D is the discriminator. The circles denote loss functions.

We view the (visual) explanation of the black-box as a generative process that produces an input for the black-box that slightly perturbs current prediction ($f(\mathbf{x}) + \delta$) while remaining plausible and realistic. By repeating this process towards each end of the binary classification spectrum, we can traverse the prediction space from one end to the other and exaggerate the underlying effect. We conceptualize the traversal from one side of the decision boundary to the other as walking across a data manifold, $\mathcal{M}_x$. We assume the walk has a fixed step size and each step of the walk makes δ change to the posterior probability of the the classifier, f. Since the output of f is bounded between $[0, 1]$, we can take at-most $\lfloor \frac{1}{\delta} \rfloor$ steps. Each positive (negative) step increases (decreases) the posterior probability of the previous step. We assume that there is a low-dimensional embedding space ($\mathcal{M}_z$) that encodes the walk. An encoder, $E : \mathcal{M}_x \to \mathcal{M}_z$, maps an input, $\mathbf{x}$, from the data manifold, $\mathcal{M}_x$, to the embedding space. A generator, $G : \mathcal{M}_z \to \mathcal{M}_y$, takes both the embedding coordinate and the number of steps and maps it back to the data manifold (see Figure1).

We use $\mathcal{I}_f(\cdot,\cdot)$ to denote the explainer function. Formally, $\mathcal{I}_f(\mathbf{x},\delta) : (\mathcal{X},\mathbb{R}) \rightarrow \mathcal{X}$ is a function that takes two arguments: a query image $\mathbf{x}$ and the desired perturbation δ. This function generates a perturbed image which is then passed through function f. The difference between the outputs of f given the original image and the perturbed image should be the desired change *i.e.*, $f(\mathbf{x}_\delta) - f(\mathbf{x}) = \delta$. We use $\mathbf{x}_\delta$ to denote $\mathcal{I}_f(\mathbf{x},\delta)$. This formulation enables us to use δ as a knob to *exaggerate* the visual explanations of the query sample while it is crossing the decision boundary given by function f. Our proposed interpretability function $\mathcal{I}_f$ should satisfy the following properties:

1. **Data Consistency:** perturbed samples generated by $\mathcal{I}_f$ should lie on the data manifold, $\mathcal{M}_x$, to be consistent with real data. In other words, the generated samples should look realistic when compared to other samples.
2. **Compatibility with f:** changing the second argument in $\mathcal{I}_f(\mathbf{x},\cdot)$ should produce the desired outcome from classifier f, *i.e.*, $f(\mathcal{I}_f(\mathbf{x},\delta)) \approx f(\mathbf{x}) + \delta$.
3. **Self Consistency:** Applying reverse perturbation should bring $\mathbf{x}$ back to its original form *i.e.*, $\mathcal{I}_f(\mathcal{I}_f(\mathbf{x},\delta),-\delta) = \mathbf{x}$. Also, applying setting δ to zero should return the query, *i.e.*, $\mathcal{I}_f(\mathbf{x},0) = \mathbf{x}$.

Each criterion is enforced via a loss function which are discussed in the following sections.

2.1 Data Consistency

We adopt the Generative Adversarial Networks (GANs) framework for our model (Goodfellow et al., 2014). The GANs implicitly model the underlying data distribution by setting up a min-max game between generative (G) and discriminative (D) networks:

$$\mathcal{L}_{\text{GAN}}(D,G) = \mathbb{E}_{\mathbf{x},c\sim P(\mathbf{x})}\big[\log\big(D(\mathbf{x})\big)\big] + \mathbb{E}_{\mathbf{z}\sim P_\mathbf{z}}\big[\log\big(1 - D(G(\mathbf{z}))\big)\big],$$

where $\mathbf{z}$ and $P_\mathbf{z}$ are the noise distribution and the corresponding canonical distribution. There has been significant progress toward improving GANs stability as well as sample quality (Brock et al., 2019; Karras et al., 2019). The advantage of GANs is that they produce realistic-looking samples without an explicit likelihood assumption about the underlying probability distribution. This property is appealing for our application.

Furthermore, we need to provide the desired amount of perturbation to the black-box, f. Hence, we use a Conditional GAN (cGAN) that allows the incorporation of a context as a condition to the GAN (Mirza & Osindero, 2014; Miyato & Koyama, 2018). To define the condition, we fix the step size, δ, and descritize the walk which effectively cuts the posterior probability range of the predictor (*i.e.*, $[0,1]$) into $\lfloor\frac{1}{\delta}\rfloor$ equally-sized bins. Hence, one can view the perturbation from $f(\mathbf{x})$ to $f(\mathbf{x})+\delta$ as changing the bin index from the current value $c_f(\mathbf{x},0)$ to $c_f(\mathbf{x},\delta)$ where $c_f(\mathbf{x},\delta)$ returns the bin index of $f(\mathbf{x})+\delta$. We use $c_f(\mathbf{x},\delta)$ as a condition to the cGAN.

The cGAN optimizes the following loss function:

$$\mathcal{L}_{\text{cGAN}}(D,G) = \mathbb{E}_{\mathbf{x},c\sim P(\mathbf{x},c)}\big[\log\big(D(\mathbf{x},c)\big)\big] + \mathbb{E}_{\mathbf{z}\sim P_\mathbf{z},c\sim P_c}\big[\log\big(1 - D(G(\mathbf{z},c),c)\big)\big], \tag{1}$$

where c denotes a condition. Instead of generating random samples from $P_\mathbf{z}$, we use the output of an encoder, $E(\mathbf{x})$, as input to the generator. Finally, the explainer function is defined as:

$$\mathcal{I}_f(\mathbf{x},\delta) = G(E(\mathbf{x}), c_f(\mathbf{x},\delta)). \tag{2}$$

Our architecture is based on Projection GAN (Miyato & Koyama, 2018), a modification of cGAN. An advantage of the Projection GAN is that it scales well with the number of classes allowing $\delta \rightarrow 0$. The Projection GAN imposes the following structure on the discriminator loss function:

$$\mathcal{L}_{\text{cGAN}}(D,\hat{G})(\mathbf{x},\mathbf{c}) = \log\frac{p_{\text{data}}(\mathbf{c}|\mathbf{x})}{q(\mathbf{c}|\mathbf{x})} + \log\frac{p_{\text{data}}(\mathbf{x})}{q(\mathbf{x})} := r(c|\mathbf{x}) + \psi(\boldsymbol{\phi}(\hat{G}(\mathbf{z}))), \tag{3}$$

where $\mathcal{L}_{\text{cGAN}}(D,\hat{G})$ indicates the loss function in Eq. 1 when $\hat{G}$ is fixed, $\boldsymbol{\phi}(\cdot)$ and $\psi(\cdot)$ are networks producing vector (feature) and scalar outputs respectively. The $r(c|\mathbf{x})$ is a conditional ratio function which will be discussed in Section 2.2.

2.2 Compatibility with the black box

In our model, the condition c is an ordered variable *i.e.*, $c_f(\mathbf{x}, \delta_1) < c_f(\mathbf{x}, \delta_2)$ when $\delta_1 < \delta_2$. Therefore, we adapt the first term in Eq. 3 to account for ordinal multi-class regression by transforming $c_f(\mathbf{x}, \delta)$ into $\lfloor \frac{1}{\delta} \rfloor - 1$ binary classification terms (Frank & Hall, 2001):

$$r(c = k|\mathbf{x}) := \sum_{i<k} \mathbf{v}_i^T \boldsymbol{\phi}(\mathbf{x}), \tag{4}$$

where $\boldsymbol{\phi}(\cdot)$ is the feature network in Eq. 3 and $\mathbf{v}_i$'s are parameters. We also need to ensure that plugging $\mathbf{x}_\delta$ into $f(\cdot)$ yields $f(\mathbf{x}) + \delta$ (*i.e.*, compatible with f). This condition is enforce by a KullbackLeibler (KL) divergence loss term. Adding the KL loss and the conditional ratio function we arrive at the following loss:

$$\mathcal{L}_f(D, G) := r(c|\mathbf{x}) + D_{\mathrm{KL}}\left(f(\mathbf{x}) + \delta \| f(\mathcal{I}_f(\mathbf{x}, \delta))\right).$$

While the first term is a function of both G and D, the second term influences only the generator G.

2.3 Self Consistency

We use a reconstruction loss term to enforce encoder-decoder consistency and satisfy the identity constraint of $\mathbf{x} = \mathcal{I}_f(\mathbf{x}, 0)$,

$$\mathcal{L}_{\text{rec}}(G) = ||\mathbf{x} - G\left(E(\mathbf{x}), c_f(\mathbf{x}, 0)\right)||_1, \tag{5}$$

We also require that the perturbation is reversible (*i.e.*, $\mathcal{I}_f(\mathcal{I}_f(\mathbf{x}, \delta), -\delta) = \mathbf{x}$). We use a cycle-consistency (Zhu et al., 2017) loss to reconstruct the input from its corresponding perturbed image,

$$\mathcal{L}_{\text{cyc}}(G) = ||\mathbf{x} - G(E(\mathbf{x}_\delta), c_f(\mathbf{x}, 0))||_1. \tag{6}$$

Note that the conditions for the generators in Eq. 5 and 6 are the same. However, in the former, we are reconstructing the input $\mathbf{x}$ from its latent space, but in the latter, we perturb $\mathbf{x}_\delta$ from the bin index $c_f(\mathbf{x}, \delta)$ back to original bin index $c_f(\mathbf{x}, 0)$.

2.4 Objective Functions

We adapted the hinge version of the adversarial loss for $\mathcal{L}_{\text{cGAN}}(G, D)$.

$$\begin{aligned} \mathcal{L}_{\text{cGAN}}(D) = &-\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\left[\min(0, -1 + D(\mathbf{x}, c_f(\mathbf{x}, 0)))\right] \\ &- \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}, c_f(\mathbf{x},\delta)\in[0,\frac{1}{\delta}]}\left[\min(0, -1 - D(G(E(\mathbf{x}), c_f(\mathbf{x}, \delta)), c_f(\mathbf{x}, \delta)))\right] \end{aligned} \tag{7}$$

$$\mathcal{L}_{\text{cGAN}}(G, E) = -\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}, c_f(\mathbf{x},\delta)\in[0,\frac{1}{\delta}]}\left[D(G(E(\mathbf{x}), c_f(\mathbf{x}, \delta)), c_f(\mathbf{x}, \delta))\right] \tag{8}$$

The overall objective function is

$$\min_{E,G} \max_{D} \lambda_{\text{cGAN}} \mathcal{L}_{\text{cGAN}}(D, G) + \lambda_f \mathcal{L}_f(D, G) + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}(G) + \lambda_{\text{cyc}} \mathcal{L}_{\text{cyc}}(G) \tag{9}$$

where $\lambda_{\text{cGAN}}, \lambda_f, \lambda_{\text{rec}}, \lambda_{\text{cyc}}$ are the hyper-parameters that balance the importance of the loss terms.

3 Related Work

Our work broadly relates to literature in interpretation methods that are designed to provide a visual explanation of the decisions made by a black-box function f, for a given query sample $\mathbf{x}$.

Perturbation-based methods: These methods provide interpretation by showing what minimal changes are required in $\mathbf{x}$ to induce a desirable output of f. Some methods employed image manipulation via the removal of image patches (Zhou et al., 2014) or the occlusion of image regions (Zhou et al., 2014) to change the classification score. Recently, the use of influence function, as proposed by (Koh & Liang, 2017) are applied as a form of data perturbation to modify a classifier's response. The authors in (Fong & Vedaldi, 2017) proposed the use of optimal perturbation, defined as removing the smallest possible image region in $\mathbf{x}$ that results in the maximum drop in classification score. In another approach, (Chang et al., 2019) proposed a generative process to find and fill the image regions that correspond to the largest change in the decision output of a classifier. To switch the decision of a classifier, (Goyal et al., 2019) suggested generating counterfactuals by replacing the regions of $\mathbf{x}$ with patches from images with a different class label. All of the aforementioned works perform pixel- or patch-level manipulation to $\mathbf{x}$, which may not result in natural-looking images. In contrast, our model enforces that the perturbed data be consistent with the unperturbed data to ensure that the perturbation is plausible. Furthermore, our method can be applied to general data and is not restricted to the imaging domain.

Saliency map-based methods: Saliency maps explains the decision of f on $\mathbf{x}$ by highlighting the relevant regions of x. Some earlier work in this direction(Simonyan et al., 2013; Springenberg et al., 2014; Bach et al., 2015) focuses on computing the gradient of the target class with respect to $\mathbf{x}$ and considers the image regions with large gradients as most informative. Building on this work, the class activation map (CAM) (Zhou et al., 2016) and its generalized version Grad-CAM Selvaraju et al. (2017) and other variants such as LPR (Bach et al., 2015) use a linear or non-linear combination of the activation layers to derive relevance score for every pixel in an image. These gradient-based methods are not model-agnostic and require access to intermediate layers. Recently, Adebayo et al. (2018) have shown that some saliency methods are independent both of the model and of the data generating process. We used their propose evaluation to validate our interpretation model. The saliency maps are also prone to adversarial attacks as shown by Ghorbani et al. (2019) and Kindermans et al. (2017). Furthermore, if the causal effect of a class is distributed across an image, which is the case in radiology images, the saliency approaches highlight large sections of the image, which greatly reduce the usefulness of the interpretation.

Generative explanation-based methods: These are a class of interpretation models that uses a generative process to produce visual explanations. A recently proposed contrastive explanations method (CEM) (Dhurandhar et al., 2018) generates explanations that show minimum regions in $\mathbf{x}$ which must be present/absent for a particular classification decision. In another work, (Liu et al., 2019; Joshi et al., 2019; Samangouei et al., 2018) generates explanations that highlight what features should be changed in $\mathbf{x}$ so that the classifier confidence in the prediction is strengthen (prototype) or weakened (counterfactual). Our approach is aligned with these latter lines of work, although our method and model architecture is different. Our method allows for the gradual change of the class effect, and our consistency criteria result in high-quality feasible perturbation in $\mathbf{x}$. We also rigorously evaluate our method on real medical imaging applications, in addition to the curated computer vision datasets.

4 Experiments

We set up four experiments to evaluate our method. First, we assess if our method satisfies the three criteria of the explainer function introduced in the Section 2. We report both qualitative and quantitative results. Second, we apply our method on a medical image diagnosis task. We use external domain knowledge about the disease to perform a quantitative evaluation of the explanation. Third, we train two classifiers on biased and unbiased data and examine the performance of our method in identifying the bias. While our method does not produce a saliency map, in our last experiment we use the two counterfactual samples on the boundary $[0, 1]$ to generate a saliency map and compare with the other methods.

Our experiments are conducted on the CelebA (Liu et al., 2014) and CheXpert (Irvin et al., 2019) datasets. CelebA contains 200K celebrity face images, each with forty attribute labels. We considered binary classifier trained on the "smiling" and "young" attributes. CheXpert is a medical dataset containing 224K chest x-ray images from 65K patients and has labels for fourteen radio-graphic observations. We considered Cardiomegaly as the target class for generating explanations. All images are re-sized to 128×128 before processing.

4.1 Evaluating the Criteria of the Explainer

Figure 2 reports the qualitative results on three datasets. Given a query image $\mathbf{x}$ at inference time, our model generates a series of images $\mathbf{x}_\delta$ as visual explanations, which gradually increase the posterior probability $f(\mathbf{x}_\delta)$ (top label). We show results for three prediction tasks: smiling or not-smiling, young or old, and Cardiomegaly or healthy. The values on the top of each figure report the $f(\mathbf{x}_\delta)$'s.

For Cardiomegaly, we show the outlines of the heart as well as its normalized size (values inside the parenthesis), which is indicative of the disease.

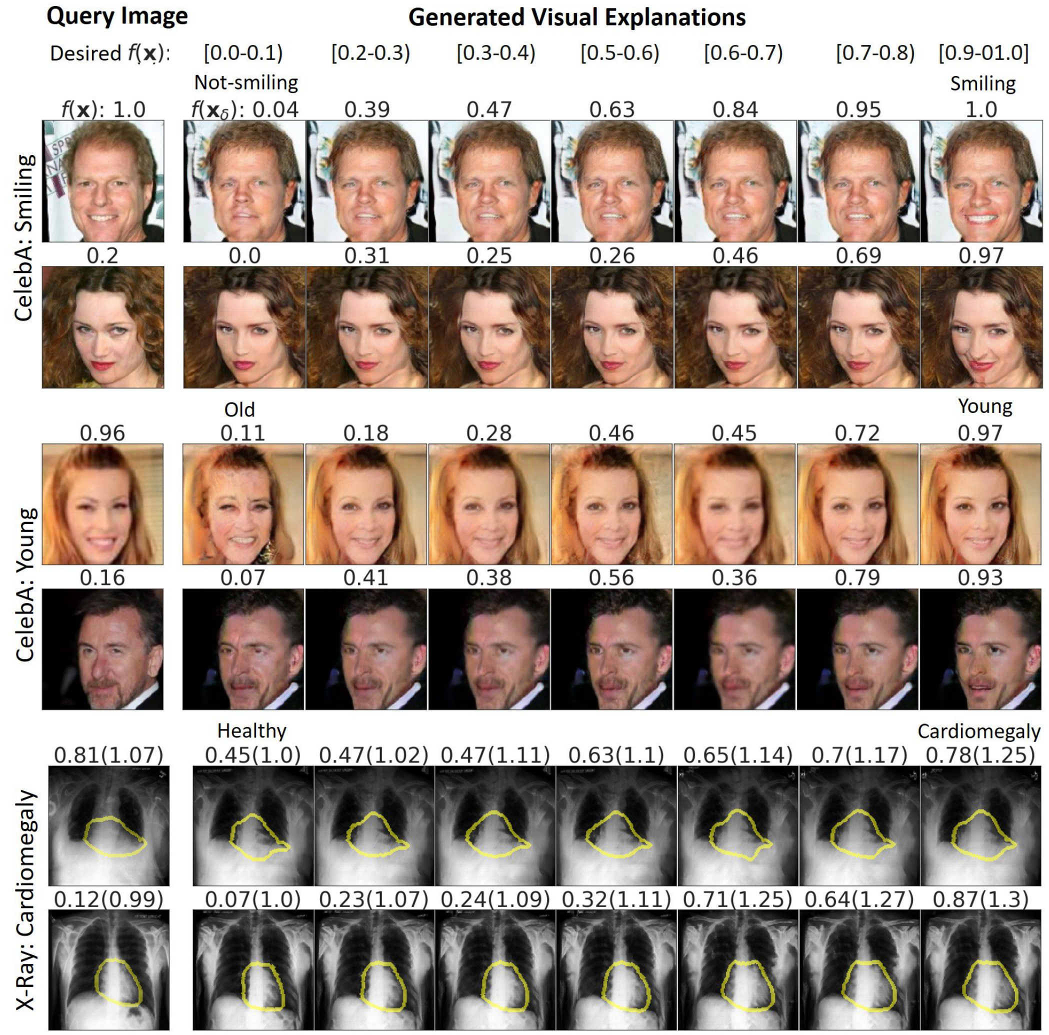


Figure 2: Visual explanations generated for three prediction tasks: smiling/not-smiling face (first two rows), young/old face (middle two rows) and Cardiomegaly/healthy chest x-ray (bottom two rows). The first column shows the query image, followed by the corresponding generated explanations. The values above each image are the output of the classifier f. For Cardiomegaly, we show the segmentation of the heart (yellow edge) and report normalized heart size (values in parenthesis), which is indicative of the disease.

Data Consistency: The generated explanations are synthesized variations of the query image. To quantitatively compare their visual quality, we consider Fréchet Inception Distance (FID) (Heusel et al., 2017). We compared our results against the counterfactual explanations produced by xGEM (Joshi et al., 2018). In the absence of the source code for the xGEM model, we implemented a baseline architecture for xGEM. We divided the real and fake (*i.e.*, generated explanations) images into two groups by partitioning on the output of the classifier f and the reported the FID for each group and the overall score. xGEM does not produce a series of progressively modified images, so we compare methods on images residing on the boundaries of the $[0, 1]$ interval. Our method significantly outperforms xGEM, producing crisper and more realistic-looking images.

Compatibility with the black-box f: To quantify whether the generation process is aligned with the desire perturbation δ, we plotted the expected outcome $f(\mathbf{x}) + \delta$ against the actual response of the classifier for the generated explanations, $f(\mathbf{x}_\delta)$. Figure 3 shows how our model performs when generating a series of explanations starting from a wide range of initial query images. The perfor-

	CelebA:Smiling		**CelebA:Young**		**Xray:Cardiomegaly**	
Target Class	xGEM	Ours	xGEM	Ours	xGEM	Ours
Present ($f(\mathbf{x}_\delta) \in [0.9, 1]$)	79.2	**52.3**	92.6	**69.0**	341.4	**55.1**
Absent ($f(\mathbf{x}_\delta) \in [0, 0.1]$)	80.6	**50.2**	95.4	**71.2**	340.6	**88.4**
Overall ($f(\mathbf{x}_\delta) \in [0, 1]$)	80.7	**23.1**	80.4	**41.6**	337.3	**50.2**

Table 1: The Fréchet Inception Distance (FID) score, measuring the quality of the generated explanations for the three prediction tasks. Lower FID corresponds to better image quality. Top (bottom) row corresponds the top (bottom) 10% of the decision interval.

mance is almost perfect for Young/Old, but less so for more challenging classification problems such as Smiling or Cardiomegaly. The plot also validates that we are producing perturb images covering the entire classification range, $[0, 1]$.

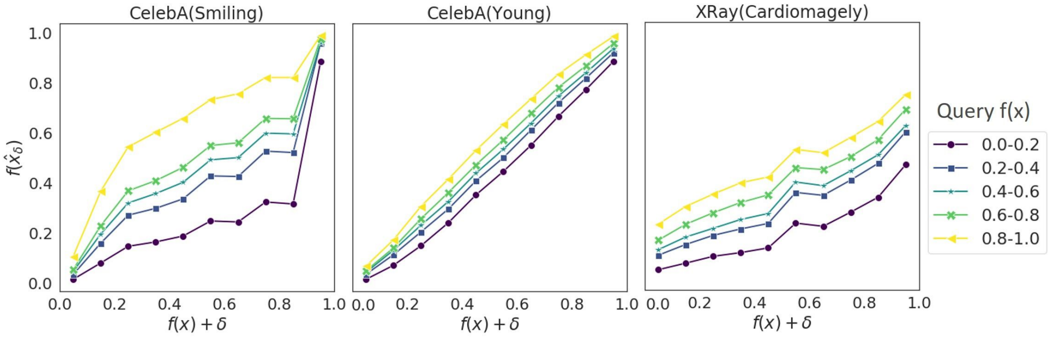


Figure 3: Plot of the expected outcome from the classifier, $f(\mathbf{x})+\delta$, against the actual response of the classifier on generated explanations, $f(\mathbf{x}_\delta)$. The monotonically increasing trend shows a positive correlation between $f(\mathbf{x}) + \delta$ and $f(\mathbf{x}_\delta)$, and thus the generated explanations are consistent with the expected condition.

Identity preservation: The generated explanations should differ only in semantic features associated with the target class, while retaining the identity of the query image. We extracted the latent embedding for real images ($E(\mathbf{x})$) and their corresponding explanations ($E(\mathbf{x}_\delta)$), for different values of δ. We calculated identity preservation accuracy as the percentage of the times, $\mathbf{x}_\delta$ is closest to the query image $\mathbf{x}$ as compared to other generated explanations

$$\forall \delta, \mathbf{x} \in \mathcal{X}, \quad ||E(\mathbf{x}) - E(\mathbf{x}_\delta)||_2 < \min_{\mathbf{m} \in \mathcal{I}_f(\mathcal{X} - \{\mathbf{x}\}, \delta)} ||E(\mathbf{x}_\delta) - E(\mathbf{m})||_2, \tag{10}$$

where, $\mathbf{m} \in \mathcal{I}_f(\mathcal{X} - \{\mathbf{x}\}, \delta))$ is the set of explanations generated for all the real images excluding the query image $\mathbf{x}$. Table 2 summarizes the results. Our method achieved high identity preservation on localized attribute "smiling", which alters a relatively small region of the face image as compare to attribute "age" which affects the entire face. Medical images like chest x-ray have very fine grain details which are difficult to preserve in the generative process of GAN. Our explainer function preserves the high level features like shape and size of the lung, but it struggles to retain the low level features like anatomy of the breast and shape of the collar bones. Also, it should be noted that both the datasets have multiple images for same person, but we ignore this information in our analysis and treat each image as a different identity.

Dataset:Target Class	**Identity Preservation Accuracy**
CelebA:Smiling	87.99%
CelebA:Young	81.63%
XRay:Cardiomegaly	27.94%

Table 2: Identity preserving performance on three prediction tasks.

4.2 Counterfactual evaluation on medical data

The term Cardiomegaly refers to an abnormal enlargement of the heart (Brakohiapa et al., 2017). We trained a UNet (Ronneberger et al., 2015) model for lung and heart segmentation using the segmentation in chest radiograph (SCR) dataset (van Ginneken et al., 2006). Next, we applied the trained UNet to our data to extract the heart masks. The masks are shown as outlines in Figure 2 with their corresponding heart size (top values in parentheses). We registered each $\mathbf{x}$ with its associated $\mathbf{x}_\delta$ and applied the resulting transformation to the heart masks of $\mathbf{x}$ to derive the heart masks for $\mathbf{x}_\delta$.

For population-level analysis, we plotted the average heart size of $\mathbf{x}_\delta$ vs the condition used for generation ($f(\mathbf{x}) + \delta$) in Figure 4 (a). The plot shows a positive correlation between the heart size and the response of the classifier $f(\mathbf{x})$, which agrees with the definition of Cardiomegaly. To better understand the results, we divided the population into two groups, the first group ($\mathbf{x}^h$; $f(\mathbf{x}^h) < 0.1$) consists of real images of healthy x-rays, while the second group ($\mathbf{x}^c$; $f(\mathbf{x}^c) > 0.9$) contains real images of abnormal x-rays positive for Cardiomegaly. For $\mathbf{x}^h$ we generated counterfactual as $\mathbf{x}^c_\delta$ such that $f(\mathbf{x}^c_\delta) > 0.9$. Similarly, counterfactuals for $\mathbf{x}^c$ are derived as $\mathbf{x}^h_\delta$ such that $f(\mathbf{x}^h_\delta) < 0.1$. In Figure 4 (b), we show the distribution of heart size in the four groups. We reported the dependent t-test statistics for paired samples $\big(\mathbf{x}^h$ and $\mathbf{x}^c_\delta\big)$, $\big(\mathbf{x}^c$ and $\mathbf{x}^h_\delta\big)$. A significant p-value $\ll 0.01$ rejected the null hypothesis (*i.e.*, that the two groups have similar distributions). We also reported the independent two-sample t-test statistics for healthy $\big(\mathbf{x}^h$ and $\mathbf{x}^h_\delta$, p-value $> 0.01\big)$ and abnormal ($\mathbf{x}^c$ and $\mathbf{x}^c_\delta$, p-value < 0.01) populations. Given higher p-values, we cannot reject the null hypothesis of identical average distributions with high confidence. Our model derived explanations successfully captured the change in heart size while generating counterfactual explanations.

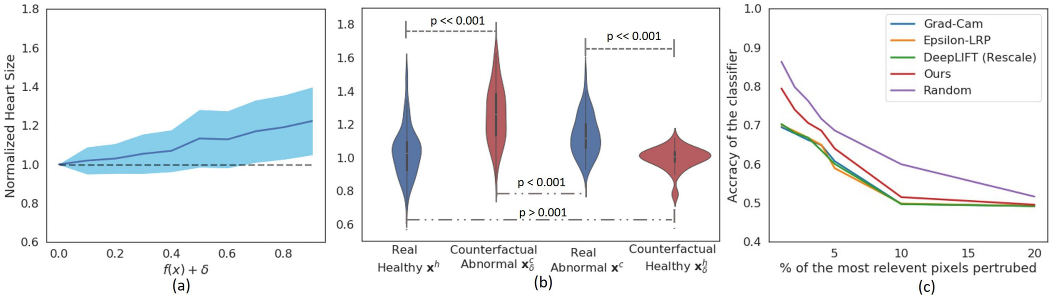


Figure 4: Cardiomegaly disease is associated with large heart size. In (a) we show the positive correlation between the heart size and the response of the classifier $f(\mathbf{x})$. (b) Comparison of the distribution of the heart size in the four groups. (c) Plot to show the drop in accuracy of the classifier as we perturb the most relevant pixels (relevance calculated from saliency map) in the image.

4.3 Saliency Map

Saliency maps show the importance of each pixel of an image in the context of classification. Our method is not designed to produce saliency maps as a continuous score for every feature of the input. We extract an approximate saliency map by quantifying the regions that changed the most when comparing explanations at the opposing ends of the classification spectrum. For each query image, we generated two visual explanations corresponding to the two extremes of the decision boundary $\big(f(\mathbf{x}_\delta) = 0$ and $f(\mathbf{x}_\delta) = 1\big)$. The absolute difference between these explanations is our saliency map. Figure 5 shows the saliency map obtain from our method and its comparison with popular gradient based methods. We restricted the saliency maps obtained from different methods to have positive values and normalize them to range [0,1]. Subjective, the saliency maps produced by our method are very localized and are comparable to the other methods.

We adapted the metric introduced in (Samek et al., 2016) to compare the different saliency maps. In an iterative procedure, we progressively replace a percentage of the most relevant pixels in an image (as given by the saliency map) with random values sampled from a uniform distribution. We observe the corresponding change in the classification performance as shown in Figure 4 (c). All the methods experienced a drop in the accuracy of the classifier with increase in the fraction of perturb

	Target Label	
Black-box classifier	Smiling	Not-Smiling
f_{Biased}	Male: 0.52 Female: 0.48 Overall: **0.12**	Male: 0.18 Female: 0.82 Overall: **0.35**
$f_{\text{No-biased}}$	Male: 0.48 Female: 0.52 Overall: 0.07	Male: 0.47 Female: 0.53 Overall: 0.08

Table 3: Confounding metric for biased detection. For target label "Smiling" and "Not-Smiling", the explanations are generated using condition $f(x) + \delta > 0.9$ and $f(x) + \delta < 0.1$ respectively. The Male and Female values quantifies the fraction of the generated explanations classifier as male or female, respectively by oracle classifier f_{Gender}. The overall value quantifies the fraction of the generated explanations who have different gender as compared to the query image. A small overall value shows least bias.

pixels. The saliency maps produced by our model is significantly better than random maps and are comparable to the other saliency map methods. It should be noted that, there are many ways to quantify important regions in a image, using the series of explanations generated by our method. We didn't optimize to find the best saliency map and showed results for one such method.

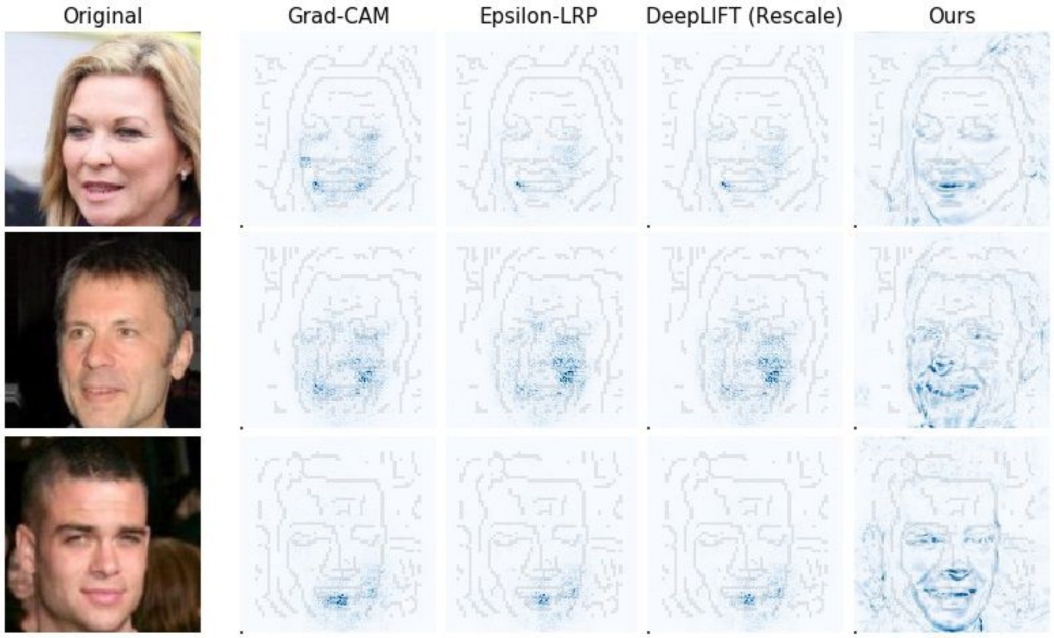


Figure 5: Our comparison with popular gradient-based saliency map producing methods on the prediction task of identifying smiling faces in CelebA dataset.

4.4 Bias Detection

Our model can discover confounding bias in the data used for training the black-box classifier. Confounding bias provides an alternative explanation for an association between the data and the target label. For example, a classifier trained to predict the presence of a disease may make decisions based on hidden attributes like gender, race or age. In a simulated experiment, we trained two classifiers to identify smiling vs not-smiling images in the CelebA dataset. The first classifier f_{Biased} is trained on a biased dataset, confounded with gender such that all smiling images are of male faces. The second classifier $f_{\text{No-biased}}$ is trained on an unbiased dataset, with data uniformly distributed with respect

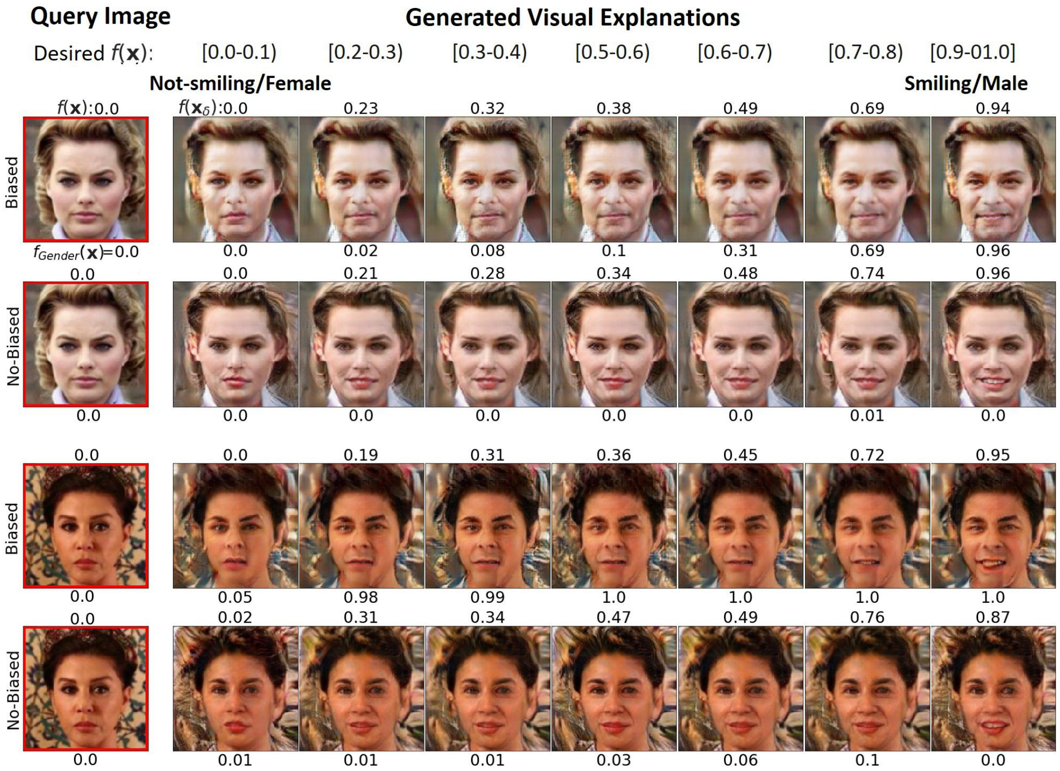


Figure 6: The visual explanations for two classifiers, both trained to classify "Smiling" attribute on CelebA dataset. For each example, the top row shows results from "Biased" classifier whose data distribution is confounded with "Gender". The bottom row shows explanations from "No-Biased" classifier with uniform data distribution w.r.t gender. The top label indicates output of the classifier and the bottom label is the output of an oracle classifier for the con-founding attribute gender. The visual explanations for the "Biased" classifier changes the gender as it adds smile on the face.

to gender. Note that we evaluate both the classifiers on the same validation set. Additionally, we assume access to a proxy Oracle classifier f_{Gender} that perfectly classifies the confounding attribute *i.e.*, gender. In Figure 6 we compare the explanations generated for the two classifiers. The visual explanations for the "Biased classifier changes the gender as it increases the amount of "smile. We adapted the confounding metric proposed in Joshi et al. (2018) to summarize our results in Table 3. Given the data $\mathcal{D} = \{(\mathbf{x}_i, y_i, a_i), \mathbf{x}_i \in \mathcal{X}, y_i, a_i \in \mathcal{Y}\}$, we quantify that a classifier is confounded by an attribute a if the generated explanation $\hat{x}_\delta$ has a different attribute a, as compared to query image $\mathbf{x}$, when processed through the Oracle classifier f_{Gender}. The metric is formally defined as $\mathbb{E}_{\mathcal{D}}[1(g^*(\mathbf{x}_\delta) \neq a)]/|\mathcal{D}|$. For a biased classifier, the Oracle function predicted female class for majority of the images (0.82) while the un-biased classifier is consistent with the true distribution of the validation set with respect to gender. Thus, the fraction of generated explanations that changed the confounding attribute "gender' (0.35 vs 0.08) was found to be high for the biased classifier

5 Conclusion

In this paper, we proposed a novel interpretation method which explains the decision of a black-box classifier by producing natural looking, gradual perturbations of the query image, resulting in an equivalent change in the output of the classifier. We evaluated our model on two very different datasets, including a medical imaging dataset. Our model produces high quality explanations while preserving the identity of the query image. Our analysis shows that our explanations are consistent with the definition of the target disease, without explicitly using that information. Our method can also be used to generate a saliency map in a model agnostic setting. In addition to the interpretability advantages, our proposed method can also identify plausible confounding biases in a classifier.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.

Sebastian Bach, Alexander Binder, Grgoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Edmund Kwadwo Kwakye Brakohiapa, Benard Ohene Botwe, Benjamin Dabo Sarkodie, Eric Kwesi Ofori, and Jerry Coleman. Radiographic determination of cardiomegaly using cardiothoracic ratio and transverse cardiac diameter: can one size fit all? Part one. *The Pan African medical journal*, 27:201, 2017. ISSN 1937-8688. doi: 10.11604/pamj.2017.27.201.12017. URL http://www.ncbi.nlm.nih.gov/pubmed/28904726http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5579422.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. 2018.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining Image Classifiers by Counterfactual Generation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1MXz20cYQ.

Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. Assessing and addressing algorithmic bias in practice. *interactions*, 25(6):58–63, 2018.

Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pp. 6967–6976, 2017.

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 *Advances in Neural Information Processing Systems*, pp. 592–603, 2018.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Ruth C. Fong and Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. ISBN 9781538610329. doi: 10.1109/ICCV.2017.371.

Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *European Conference on Machine Learning*, pp. 145–156, 2001.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019.

Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pp. 227–236. ACM, 2008.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual Visual Explanations. *CoRR*, abs/1904.07451, 2019. URL http://arxiv.org/abs/1904.07451.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. 1 2019. URL http://arxiv.org/abs/1901.07031.

Shalmali Joshi, Oluwasanmi Koyejo, Been Kim, and Joydeep Ghosh. xGEMs: Generating Examplars to Explain Black-Box Models. *arXiv preprint arXiv:1806.08867*, 2018.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards Realistic Individual Recourse and Actionable Explanations
}in Black-Box Decision Making Systems. *CoRR*, abs/1907.09615, 2019. URL http://arxiv.org/abs/1907.09615.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Been Kim. *Interactive and interpretable machine learning models for human machine collaboration*. PhD thesis, Massachusetts Institute of Technology, 2015.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *arXiv preprint arXiv:1711.00867*, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *34th International Conference on Machine Learning, ICML 2017*, 2017. ISBN 9781510855144.

Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative Counterfactual Introspection for Explainable Deep Learning. *arXiv preprint arXiv:1907.03077*, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. 11 2014. URL http://arxiv.org/abs/1411.7766.

Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *CoRR*, abs/1411.1784, 2014. URL http://arxiv.org/abs/1411.1784.

Takeru Miyato and Masanori Koyama. cGANs with Projection Discriminator. *CoRR*, abs/1802.05637, 2018. URL http://arxiv.org/abs/1802.05637.

Christoph Molnar. Interpretable machine learning. *A Guide for Making Black Box Models Explainable*, 7, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

Pouya Samangouei, Ardavan Saeedi, Liam Nakagawa, and Nathan Silberman. Explaingan: Model explanation via decision boundary crossing transformations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 666–681, 2018.

Wojciech Samek, Alexander Binder, Grgoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

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*, pp. 618–626, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A Riedmiller. Striving for Simplicity: The All Convolutional Net. *CoRR*, abs/1412.6806, 2014.

Ryan Turner. A model explanation system. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2016.

B van Ginneken, M B Stegmann, and M Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical Image Analysis*, 10(1):19–40, 2006.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial
}Networks. *CoRR*, abs/1703.10593, 2017. URL http://arxiv.org/abs/1703.10593.