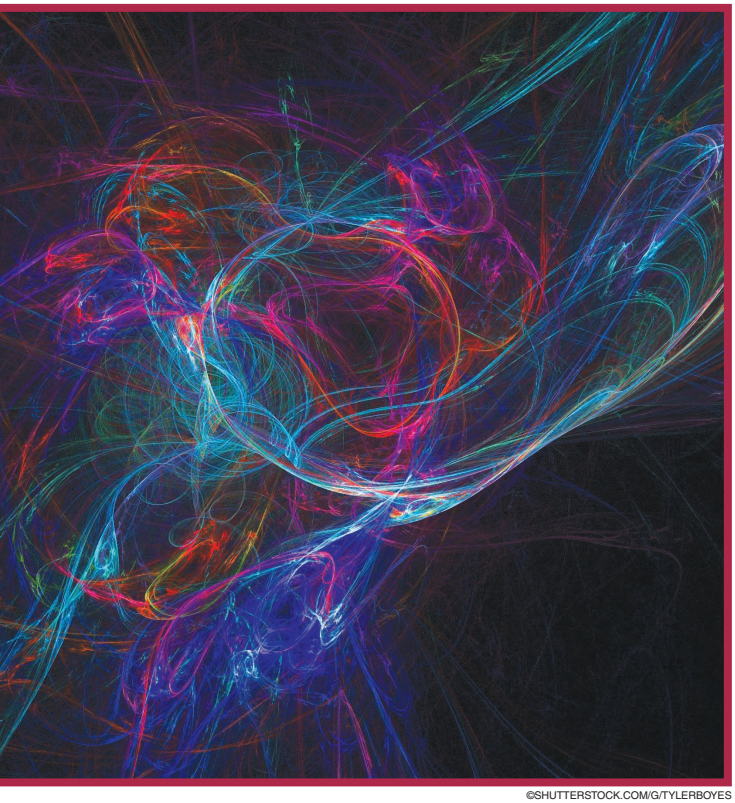
Ghassan AlRegib and Mohit Prabhushankar

Explanatory Paradigms in Neural Networks

Towards relevant and contextual explanations



©SHUTTERSTOCK.COM/G/TYLERBOYES

Digital Object Identifier 10.1109/MSP.2022.3163871
Date of current version: 28 June 2022

In this article, we present a leap-forward expansion to the study of explainability in neural networks by considering explanations as answers to abstract reasoning-based questions. With P as the prediction from a neural network, these questions are "Why *P*?", "What if not *P*?", and "Why *P*, rather than *Q*?" for a given contrast prediction *Q*. The answers to these questions are observed correlations, counterfactuals, and contrastive explanations, respectively. Together, these explanations constitute the abductive reasoning scheme. The term observed refers to the specific case of posthoc explainability when an explanatory technique explains the decision *P* after a trained neural network has made the decision. The primary advantage of viewing explanations through the lens of abductive reasoning-based questions is that explanations can be used as reasons while making decisions. The posthoc field of explainability, which previously justified decisions, becomes active by being involved in the decision-making process and providing limited but relevant and contextual interventions. The contributions of this article are 1) realizing explanations as reasoning paradigms, 2) providing a probabilistic definition of observed explanations and their completeness, 3) creating a taxonomy for evaluation of explanations, and 4) positioning gradient-based complete explainability's replicability and reproducibility across multiple applications and data modalities, and 5) code repositories, which are publicly available at https://github.com/olivesgatech/Explanatory-Paradigms.

Introduction

The generalizability of machine learning has fostered its ubiquitous use across diverse applications within artificial intelligence (AI). AI takes an inferential and interpretive role in safety-critical fields of medicine, remote sensing, and bioinformatics [1], among others. For example, AI algorithms detect cardiovascular risk factors in humans through retinal fundus images [2]. In [3], the authors propose an AI-based intervention system that identifies early red flags in law enforcement to prevent police-public adverse events. In such sensitive applications, it is apparent that the role of inference in an AI system is not a binary conclusion of whether risks exist. Rather,

AI inference must provide an explanation about the risks involved in the decision, thereby allowing humans to assess and act on the risks. Consequently, Poplin et al. [2] provide a visual explanatory map that shows why and where an AI model detects cardiovascular risk symptoms. A number of works exist to track the larger concept of interpretability in AI. The authors in [4] provide a comprehensive overview of AI interpretability and its larger denomination with related fields of privacy, transparency, safety, explainability, and ethics. In this article, we specifically focus on explainability of visual data for a specific modality of machine learning, namely, neural networks.

The generalization and task-independent modeling capabilities of neural networks have resulted in their widespread applicability. *Generalization* refers to a neural network's ability to regularize to unseen data. Effective generalization is required for large-scale learning. *Task-independent modeling* refers to transference of low-level features between tasks and networks. For instance, pretrained neural networks trained on natural images are used on computed seismic images [5]. These two capabilities, coupled with hardware improvements, have allowed neural networks to achieve state-of-the-art performance on a number of tasks. Specifically in image classification, neural networks have surpassed top-five human accuracy of 94.9% on the ImageNet data set [6]. Although gaining such high-performance accuracy is an achievement, it is equally important to understand the reasons that lead to such high accuracy rates. Thus, it is imperative to study explainability in neural networks to better understand the underlying science behind the phenomenon. However, explainability in such networks is challenging as neural networks do not conform to the traditional notion of AI interpretability. An example of AI interpretability is through decision trees [7]. The idea that neural networks should obtain an explainability similar to that of decision trees is not unreasonable. Decision trees are nonlinear models that create local binary partitions. This is not dissimilar to neural net neurons with rectified linear unit (ReLU) activation functions that either propagate or suppress features. Providing such an explanation requires a complete understanding of features at each node. Although decision trees are designed using disentangled and interpretable features, neural networks learn such features along with feature interactions. This challenge has recently been compounded by the scale of neural networks. A generative pretrained transformer-3 natural language processing model [8] consists of 175 billion parameters. Parsing all the parameters and decisions within such models is impractical. Hence, explainability in neural networks is abstract. This abstract definition of explainability in neural networks is motivated by the broad definition of explanations themselves.

Explanations are defined as a set of rationales used to understand the reasons behind a decision [9]. If the decision is based on visual characteristics within the data, the decision-making reasons are visual explanations. Consider the spoonbill image in Figure 1. A trained neural network correctly recognizes it as a spoonbill. Consider the form of explanations that are acceptable for such a decision. In Figure 1, we present three such toy explanations, all of which allude to the general goals of explainability. All three explanations are abstract answers to specific questions. When a child is interested in knowing "Why Spoonbill?", the network's explanation highlights the body and head of the bird leading to that decision. This is an abstract explanation that requires humans to take the extra step of knowing that the highlighted regions depict the spoonbill's flat beak and its pink and round body. For the neural network, these characteristics determine the

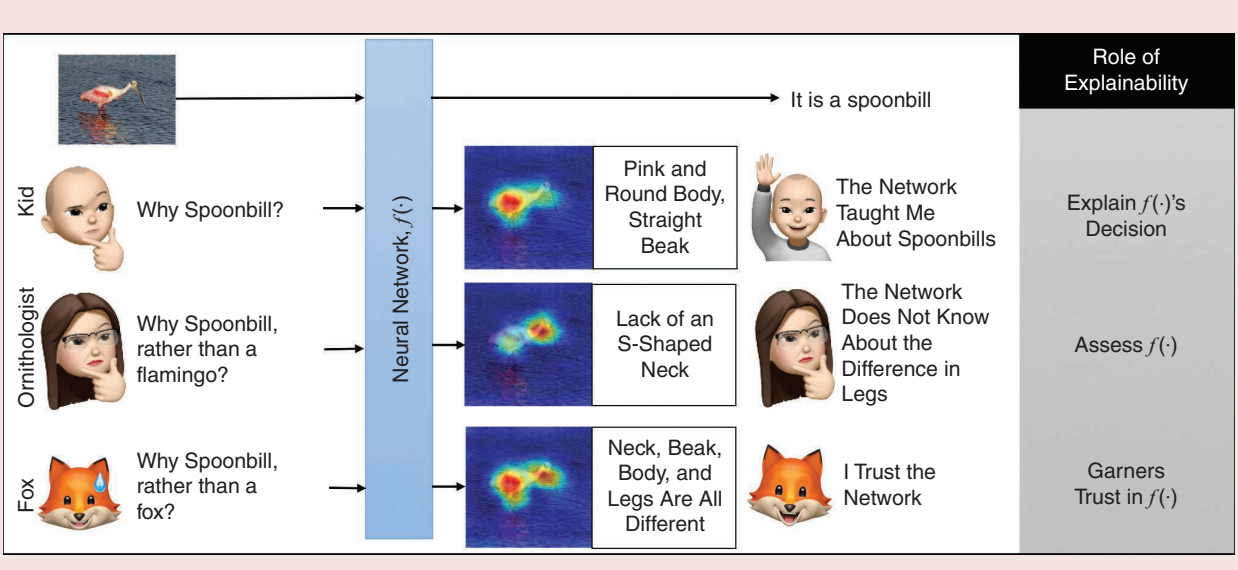


FIGURE 1. The role of explainability demonstrated through targeted toy examples. Note that $f(\cdot)$ is not trained to answer abstract questions. The explanations provide additional insights beyond the trained solution. (Source: [42] (spoonbill).]

class of the given bird. Note that these are not causal factors. Flamingos have the same body shape and color that the explanation highlights. Consider the scenario when an ornithologist, armed with the knowledge that flamingos and spoonbills share the same body shape and color, asks the question, "Why spoonbill, rather than a flamingo?" Such a targeted and contextualized question elicits the explanation that the two birds differ in the shape of their necks. Specifically, the lack of an S-shaped neck in a spoonbill is highlighted as to why it is not a flamingo. The relevance of the lack of an S-shaped neck is accentuated by the question, "Why spoonbill, rather than a flamingo?" Such an explanation can be used to assess the network. In the last toy example, the network answers the question, "Why spoonbill, rather than a Fox?" by highlighting all the body parts of the bird. Note that the network is trained only to classify birds and has not explicitly been trained to answer these questions. Hence, all these intelligible and abstract explanations together serve to garner trust in the network and its inference capabilities.

The central thesis of this article is to motivate and categorize existing abstract explanatory techniques based on specific questions. We consider three types of questions, namely, observed correlations "Why P?", observed counterfactual "What if?", and observed contrastive "Why P, rather than Q?" Here, P is any prediction from a neural network, like a spoonbill, and Q is any contrast class, like a flamingo or a fox. The answers to these questions are observed in the sense that the decision has already been made. Together, these three questions provide complementary explanations that we term *explanatory paradigms*. Their complementary nature allows us to define complete explanations as justifications for any and all posed contextually relevant questions. For the observed counterfactual and contrastive paradigms, there is an additional intervention that allows for their respective questions. These explanations are not *interventionist* from a causal perspective [10]. *Interventions* refer to manipulations in data that are designed to probe a network's class causality and construct counterfactuals. In this article, the interventions considered are based on a network's notion of classes, through gradients. By considering such interventions, we elicit model behaviors that provide responses to the ornithologist and the fox from Figure 1 without the need for engineered interventions.

Background

We first discuss requisite neural network preliminaries. Although the focus is on classification networks, the preliminary discussion extends to other discriminative and generative networks as well. This is followed by existing as well as our own definitions of explainability.

Neural network preliminaries

Classification neural networks learn nonlinear transformations to obtain discriminative representation spaces on any given data. The inherent mechanisms behind the discriminative process can be formulated as follows. Let $f(\cdot)$ be a neural network trained to distinguish among N classes. If x is any input to the network, the output logits in a classification network are given by

$$\hat{y} = f(x), \forall \hat{y} \in \Re^{N\times 1}, \tag{1}$$

where $\hat{y}$ is an $(N \times 1)$ vector. The predicted class of x, given by P, is the index of the maxima of $\hat{y}$, i.e.,

$$P = \operatorname{argmax} \hat{y}, P \in [1, N]. \tag{2}$$

During training, the network is penalized for predicting the wrong class. The penalty is calculated using an empirical loss function, $J(P, y)$, where y is the true label for input x. This loss is backpropagated through the network. The network parameters are updated based on gradients $\partial J(P, y)/\partial\theta$, where θ are the network parameters, namely, its weights and biases. Mathematically, this update is given by

$$\theta' = \theta - \frac{\partial J(P, y)}{\partial \theta}, \tag{3}$$

where θ' are the parameters after the update. A full description of backpropagation and the associated math is presented in [11]. The gradients can be backpropagated all the way back to the input x to obtain $\partial J(P, y)/\partial x$ or to the activations at some layer L, given by A_L, to obtain $\partial J(P, y)/\partial A_L$. These quantities represent *sensitivity*, which is defined as a measure of the change in the prediction loss given a small perturbation either in x or A_L. The parameter gradients and sensitivity are widely used as explanatory features among a number of explanatory methods, including [12]–[18].

Interventions, causality, and posthoc explanations

The interventions in data are manipulations designed to test for causal factors [10]. The authors in [19] use the presence of interventions to motivate two types of data, namely, observational and interventionist. The interventionist data are specifically engineered to test for causal variables. They employ the *do* operator and are generally given as $\mathbb{P}(Y \mid do(x))$ for data x and label Y. The *do* operator represents interventions. However, all possible interventions in data can be long, complex, and impractical [20]. Even with interventions, it is challenging to estimate whether the resulting effect is a consequence of the intervention or due to other uncontrolled interventions. Instead of interventions, posthoc explanations highlight features using a trained network that aid inference. In this article, we use the term *observed* to denote a specific case of posthoc explainability after a network has made its decision. We then construct gradient-based observed interventions to create complete observed explanations.

Explanation maps

The definition of explanatory maps in neural networks has evolved over time. The early works for explainability [21], [22] consider explanations to be processes that illustrate the workings of neural networks. The deconvolution-based explainability

technique proposed in [21] shows that at different layers and different kernels, diverse structures in the data activate to create local decisions. Deconvolution, unpooling, and rectification are used to reconstruct and visualize activations at all layers. These visualizations are analyzed based on layers and kernels to obtain insight regarding individual parameters within the network. Such a hierarchical representation is followed in [23], where the authors consider a neural network as an explanatory graph. From this methodology of obtaining explanations, one could define explanations as visualizations that require further deductions regarding the network and its outputs. We term such explanations as *indirect explanations*. An alternative definition of explainability is based on [24] and defines visual explanations as localization maps. Such a definition is based on the argument that neural networks, with no explicit training, implicitly localize the objects they recognize. We term these types of explanations *direct explanations*. Direct explanations passively highlight all the regions in an image that are used to predict *P*. Zhou et al. [24] produce a class activation map (CAM) that highlights the parts of the image that lead to the final decision. Direct explanations are passive in the sense that they retroactively justify decisions.

Recently, however, direct and indirect definitions of explanations have been insufficient to categorize some of the newer explanatory techniques [12], [13]. The authors in [12] show explanations for any class in any given image, even when objects from that class are not present in the image. Prabhushankar et al. [13] show the contrast between any two classes in the image. This is depicted in the contrast between a spoonbill and a flamingo in Figure 1. Neither the definitions of localization, nor reconstruction encompass these works. However, such works lead to explanations taking an active role in neural networks by participating in decision making or assessing their parent networks [14]. These explanations encompass all three explanatory roles sought in Figure 1. In this article, we propose a new and simple probabilistic definition for explanations that combine newer and older explainability techniques, and construct a more comprehensive paradigm.

To better understand and distinguish the definitions of explainability, consider an image x fed into a trained network $f(\cdot)$. The network is trained to distinguish x between N classes as $P \in Y$. Let $\mathcal{T}$ be the set of all features on which the network has learned to base its decision. In other words, a network infers that $x \in P$, if x has features $\mathcal{T}_p$. Hence $\mathcal{T} = [\mathcal{T}_1, \mathcal{T}_2 \ldots \mathcal{T}_M]$, where M is the total number of features that is learned by a network. Note that we are not considering $\mathcal{T}_i, i \in [1, M]$ to be the causal features. In other words, $\mathcal{T}_i \cap \mathcal{T}_j \neq 0, \forall i \neq j$. The features from multiple sets constitute $\mathcal{T}_p$ and are used to make inferences on x, i.e., $\mathcal{T}_p = \cup_{i=1}^{K_p} \mathcal{T}_i, K_p \in [1, M]$. We define an explanatory map $\mathcal{M}$ as follows.

Definition 2.1 (observed explanation)

An explanatory map is any function $\mathcal{M}(\cdot)$ that maximizes the conditional probability $\mathbb{P}(\cup_{i=1}^{K} \mathcal{T}_i \,|\, Y)$, where Y is any possible class, $K \in [1, M]$ and $\cup_{i=1}^{K} \mathcal{T}_i$ is the union of all required features $\mathcal{T}_i$ that specifically lead to Y.

In "Definition 2.1," we assume decision Y before maximizing features $\mathcal{T}_i$. However, by not specifying the value for Y, we allow for interventions in labels to create contrastive paradigms. By specifying the features as noncausal $\mathcal{T}_i$, we open the possibility of changing features to create counterfactual paradigms. This allows specific interventions in Y and $\mathcal{T}_i$. "Definition 2.1" is not causal interventionist because we neither alter the feature set $\mathcal{T}$, interventionist counterfactual, nor probe on x for interventionist causality. Our definition provides for restricted interventions for a decision P, and we term it *observed*. Our observed definition of explanation comprises both indirect and direct explanatory techniques. Direct and indirect explanations maximally highlight features that explain decision P. In other words, their goal is to maximize $\mathbb{P}(\mathcal{T}_p \,|\, P)$ for the specific instance of $Y = P$. Our general definition explains any decision Y by maximally highlighting features $\cup_{i=1}^{K} \mathcal{T}_i$. The newer explanatory techniques like [12] and [13], which do not conform to direct and indirect definitions of explanations, can be explained using "Definition 2.1." This is presented in the "Explanatory Paradigms" section.

Explanatory paradigms

Observed explanatory paradigms via reasoning

We first motivate our observed paradigms through reasoning. The primary goal of neural networks is to infer decisions. Any network $f(\cdot)$ is trained to maximize $\mathbb{P}(P \,|\, \mathcal{T}_p)$, i.e., $\mathcal{T}_p$ is the reason that leads $f(\cdot)$ to predict P. Hence, explanations are a means by which to extract these reasons. There are three ways in which reasoning can occur: deductive, inductive, and abductive [14]. Deductive reasoning occurs when networks are provided with a set of features and logic to use the given features. As both the features and the logic used to manipulate the features are learned by the network, deductive reasoning does not play a large role in neural networks. Inductive reasoning forms the core of the reasoning behind current networks. A set of parameters are learned, consisting of some underlying rules and knowledge. During inference, neural networks associate given data with the knowledge in their parameters and make their predictions with a probability. The explanations are then generated to justify such a decision. Direct and indirect explanations are products of this reasoning scheme. The third form of reasoning is abductive reasoning. The primary difference between inductive and abductive reasoning is the use of explanations. In abductive reasoning, an explanation is any hypothesis that supports a prediction. In Figure 1, all three explanations support the prediction that the given image is that of a spoonbill. The network can detect the presence of the pink and round body and straight beak to inductively infer the class of the bird as spoonbill. Or, alternatively, it can hypothesize as to why the bird cannot be a flamingo—its lack of an S-shaped neck—and then decide that the bird has to be a spoonbill. Such an inference is abductive. A number of such hypotheses can be constructed, fused, and incorporated together, then used to make a decision.

In this article, we present three specific ways of generating such hypotheses as answers to causal questions, "Why *P*?"; counterfactual questions, "What if not P?"; and contrastive questions, "Why *P*, rather than *Q*?". However, as we use observed techniques, the network provides correlation features to the causal question, and we name it *observed correlation*. In the "Probabilistic Completeness of Explanations" section, we show that these three explanatory paradigms are sufficient to generate probabilistically complete explanations. A visual depiction of the three paradigms is presented in Figure 2. Consider a Bullmastiff as the given input image. The observed correlation explanation answers, "Why Bullmastiff?" by highlighting the face of the dog. (Note that Grad-CAM [12] is used to obtain this explanation.) The question, "What if the Bullmastiff was not in the image?" is a counterfactual question. As highlighted using the technique in [12], the network would have made a decision based on the cat for the given counterfactual. The contrastive question is of the form "Why Bullmastiff, rather than boxer?" The heavy jowls of the Bullmastiff and the presence of a cat preclude the network from predicting a boxer, as highlighted by Contrast-CAM from [13].

Contrastive explanations take into account the context provided by the questionnaire, thus making explanations more relevant. Note that in all three cases, the features set $\mathcal{T}$ highlighted by the explanations are different. However, finding any feature set $\mathcal{T}$ in neural networks is challenging because features $[\mathcal{T}_i], \forall i \in [1, M]$ are not disentangled. Rather, feature extraction and inference occur simultaneously by projecting data onto network parameters. We consider $\mathcal{T}$ to be the span of network parameters or as manifolds in a high-dimensional space. This manifold is, however, different for dissimilar network architectures. Hence, each feature set $\mathcal{T}$ is tied to a specific neural network $f(\cdot)$ as $\mathcal{T}^f$. To keep the notations simple, we ignore the superscript and name the three paradigms *observed correlation*, *observed counterfactual*, and *observed contrastive explanations*. Observed, in this article, is from the perspective of a neural network.

Observed correlation explanations

Observed correlation explanations are answers to "Why *P*?" questions. *P* can be any learned class but is generally the prediction made by the network. Observed correlation explanations can be modeled using any of the indirect or direct explanations from the "Explanation Maps" section. Using "Definition 2.1," observed correlation explanations $\mathcal{M}_{\text{cu}}(\cdot)$ find the feature set $\mathcal{T}_p$ that satisfies $\mathbb{P}(\mathcal{T} = \mathcal{T}_p \,|\, Y = P)$. In Figure 3(a), we visually represent the search for observed correlation features for the example of the spoonbill image. Consider the blue manifold as the learned manifold where the image of the spoonbill is classified as a spoonbill. The set of all possible features $[\mathcal{T}_i], \forall i \in [1, M]$ exists on the blue manifold. The goal of any observed correlation explanation is to find features $\mathcal{T}_p = \cup_{i=1}^{K} \mathcal{T}_i$, which span the purple manifold within the larger blue manifold. Although the blue manifold uses all the parts of x to recognize a spoonbill, only the features in its legs and neck are sufficient for x's recognition. Hence, $\mathcal{T}_p$ are features in a spoonbill's neck and legs.

Note that the observed correlation features presented here are different from the interventionist causal factors discussed in [10], [19], and [25], even though the questions they answer are causal. These are posthoc explanations observed from a network. The observed correlation explanations take prediction *P* into consideration to retroactively extract features $\mathcal{T}_p$. Common, existing observed correlation techniques are listed in Table 1.

Observed counterfactual explanations

Counterfactual explanations are contextually relevant explanations that are of the paradigm "What if?" The explanation to "What if the Bullmastiff was not in the given image?" is shown in Figure 2. The other counterfactual questions can take the form "What if the dog had feathers?" or "What if the dog had a longer neck?" Generally, these questions are tackled by intervening within the images in [10], [19], and [26]. In the observed paradigm, we intervene in feature space $\mathcal{T}$. Conceptually, this is similar to maximizing $\mathbb{P}(do(\mathcal{T}) \,|\, Y = y)$ from the "Interventions, Causality, and Posthoc Explanations"

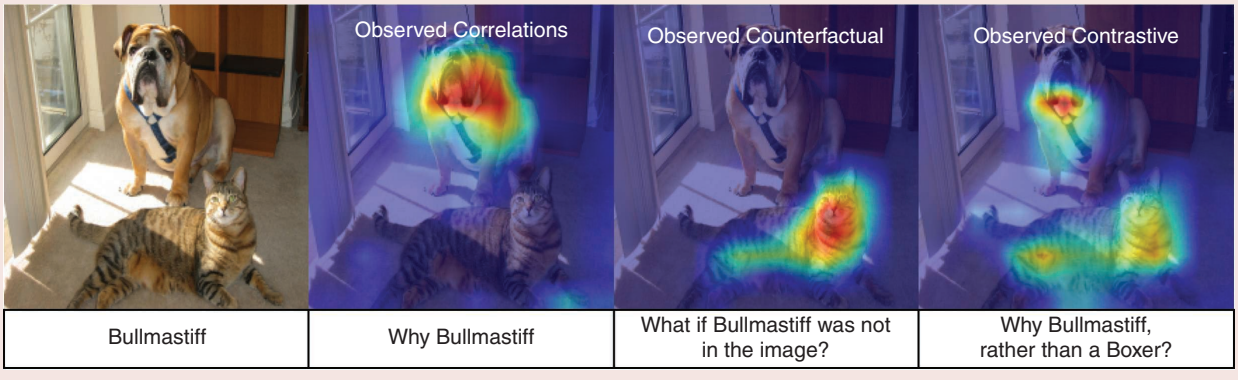


FIGURE 2. The three paradigms of explanations, each answering observed correlation, counterfactual, and contrastive questions, respectively. (Source: [42] (Bullmastiff).]

section. These *do* interventions are limited because original feature set $\mathcal{T}$ remains unchanged. Depending on the question asked, observed counterfactual explanations $\mathcal{M}_{cf}(\cdot)$ are given by the probability of $\mathcal{T}_{\hat{y}}$, which satisfy $\mathbb{P}(\mathcal{T} = \mathcal{T}_{\hat{y}} \,|\, Y = y)$. $\mathcal{T}_{\hat{y}}$ is defined similarly to observed correlation explanations, except that there is an intervention in the feature set of the label depicted by $\hat{y}$. This intervention can either be the absence of y, like the counterfactual question in Figure 2, or alterations in the data that change y, like in [26].

In Figure 3(b), we visualize three possible interventions on the spoonbill when $y = P$. The first two interventions occur on the legs and neck of the spoonbill, respectively. These intervened images, x_1 and x_2, are projected onto separate counterfactual manifolds and classified as flamingos. The last intervention occurs in the body of the flamingo but is insufficient to change its classification. Nevertheless, x_3 is projected onto a different set of features that span a new manifold. Hence, in this toy example, there are three separate $\hat{p}$ and three $\mathcal{T}_{\hat{p}}$. Note that the underlying features, $\mathcal{T} = [\mathcal{T}_1, \mathcal{T}_2 \ldots \mathcal{T}_M]$, remain the same. The union of the subsets given by $\cup_{i=1}^{K} \mathcal{T}_i$ is different. This suggests that no new features are learned, rather a new union of existing features is formed. The result shown in Figure 2 illustrates how the network can highlight features in the cat when denied features of the dog. The authors in [20] show that interventions can be long, complex, and impractical.

Observed contrastive

The third explanatory paradigm of contrastive explanations provide an intuitive resolution to the intervention complexity challenge by taking advantage of the implicit notion about individual classes learned by the network. They answer questions of the form "Why P, rather than Q?", where P is the predicted class and Q any learned contrast class. Contrastive explanations take into account the context provided by the questionnaire, thus making explanations more relevant. For a given class Q and predicted class P, contrastive explanations $\mathcal{M}_{\text{ct}}(\cdot)$ maximize probability $\mathbb{P}(\mathcal{T} = \mathcal{T}_{p,q} \,|\, Y = Q)$, where $\mathcal{T}_{p,q} = \cup_{q=1}^{K_1} \mathcal{T}_p \cap \mathcal{T}_q$. From the "Interventions, Causality, and Posthoc Explanations" section, we are intervening in the label space as $\mathbb{P}(\mathcal{T} \,|\, do(Y))$ to create an observed contrastive paradigm.

Consider the learned blue manifold in Figure 3(c), which consists of all features $\mathcal{T}$ that are used to recognize a spoonbill. Contrastive explanations require a new hypothetical manifold that produces enough interventions in $\mathcal{T}$ to predict a spoonbill as a flamingo. This occurs when the parameters of the blue manifold are changed to obtain the contrastive purple manifold. The gradients described in (3) are used to adjust parameters between the learned and contrastive manifolds in [13].

Probabilistic completeness of explanations

An advantage of defining the three observed explanatory maps and their associated paradigms in a probabilistic fashion is that they can be combined into complete explanations. The authors in [27] describe completeness in neural networks as explaining all parameters. In this article, we explain all probabilities. From (2), a neural network can have N class probabilities. An explanatory technique must answer for any one of these possible decisions in the form of correlation-based, counterfactual, and contrastive questions. Such an explanation is termed a *complete explanation*. Here, rigorous definition is provided.

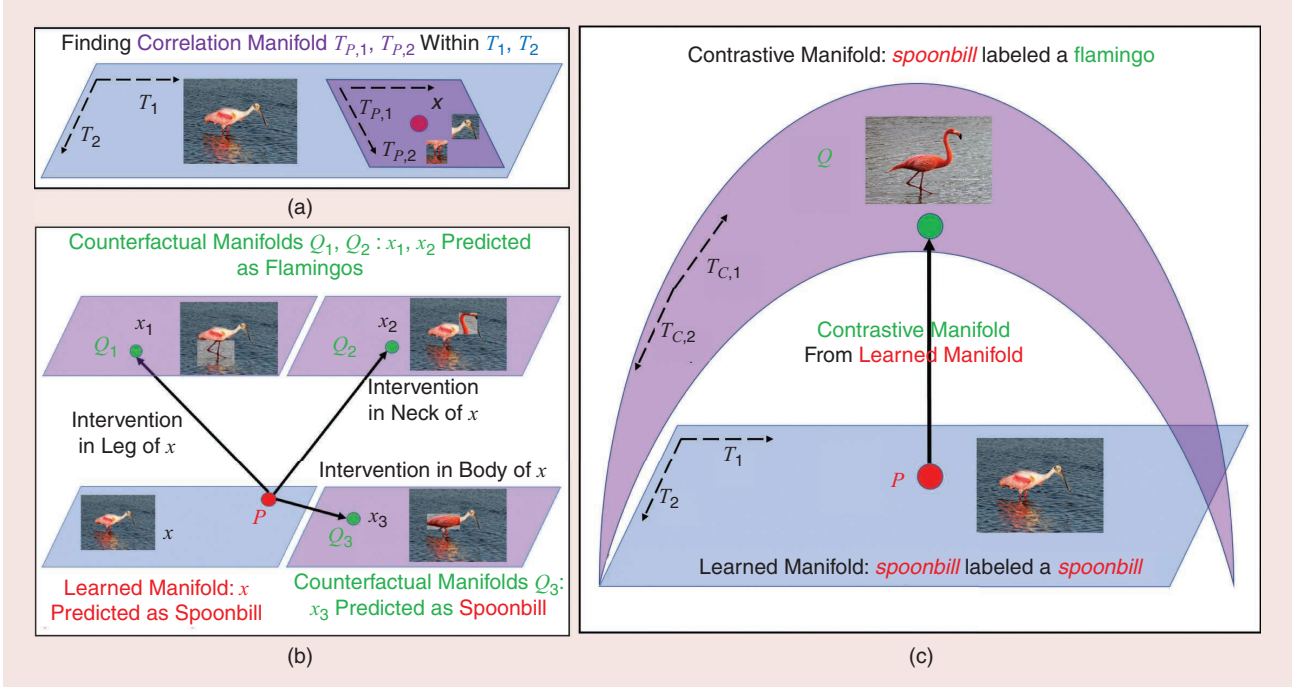


FIGURE 3. Three paradigms of explanations, and their manifolds. All requisite changes in the manifolds can be tracked using gradients from (3).

Definition 3.1 (complete explanation)

A complete explanation, $\mathcal{M}_c(\cdot)$, is a union of all functions $\mathcal{M}(\cdot)$ that simultaneously maximize conditional probabilities $\mathbb{P}(\mathcal{T} \mid Y)$ for all $Y \in [1, N]$ and all features $\mathcal{T}$, i.e., $\mathcal{M}_c(\cdot) = \cup_{j=1}^{N} \mathcal{M}_j(\cdot) = \cup_{j=1}^{N} \mathbb{P}(\cup_{i=1}^{K} \mathcal{T}_i \mid Y_j)$,
where $\mathcal{M}(\cdot), \mathcal{T}, Y$, and K are defined in "Definition 2.1." Complete explanations can simultaneously answer questions regarding any class Y. For instance, in Figure 2, given a dog and a cat image, a question of the form "Why blue jay?" is an acceptable one based on "Definition 3.1" as it probes $Y = y_{\text{Bluejay}}$, and because $f(\cdot)$ has learned the blue jay class. However, it is irrelevant seeing that there is no blue jay in the image. An alternative and relevant question of the form "Why Bullmastiff, rather than a blue jay?" can be used to probe $Y = y_{\text{Bluejay}}$. An explanation for such a question highlights the face of the dog [14]. Note that the three paradigms provide a way to probe all class probabilities in a contextually relevant fashion.

Consider a binary classifier $f(\cdot)$, with P and Q as the two possible classes. Let $\mathbb{P}(P)$ and $\mathbb{P}(Q)$ be the probabilities of $f(\cdot)$, predicting a given image x that belongs to P and Q, respectively. For a binary classifier, the sum of both probabilities adds up to one,

$$1 = \mathbb{P}(P) + \mathbb{P}(Q). \tag{4}$$

Decomposing (4) using the law of total probability, we have

$$\begin{aligned} 1 = {} & \mathbb{P}(P \mid \mathcal{T}_p)\mathbb{P}(\mathcal{T}_p) + \mathbb{P}(P \mid \mathcal{T}_p^c)\,\mathbb{P}(\mathcal{T}_p^c) + \mathbb{P}(Q \mid \mathcal{T}_q)\mathbb{P}(\mathcal{T}_q) \\ & + \mathbb{P}(Q \mid \mathcal{T}_p^c)\,\mathbb{P}(\mathcal{T}_q^c), \end{aligned} \tag{5}$$

where $\mathcal{T}_p^c$ and $\mathcal{T}_q^c$ are complements. In other words, $\mathcal{T}_p^c$ refers to the set of features not used to make decision P, and $\mathcal{T}_q^c$ are the features not used to decide Q. Using Bayes' theorem on all terms on the right-hand side, by eliminating probabilities of the features in the numerator and denominator and rearranging to obtain $\mathcal{T}_p^c$ with $\mathcal{T}_q^c$, we get

$$\begin{aligned} 1 = {} & \mathbb{P}(\mathcal{T}_p \mid P) \times \mathbb{P}(P) + \mathbb{P}(\mathcal{T}_q \mid Q) \times \mathbb{P}(Q) + \mathbb{P}(\mathcal{T}_p^c \mid P) \\ & \times \mathbb{P}(P) + \mathbb{P}(\mathcal{T}_q^c \mid Q) \times \mathbb{P}(Q). \end{aligned} \tag{6}$$

$\mathcal{T}_p^c$ is analogous to the form "What if the Bullmastiff was not in the image?" in Figure 2, where the network makes a decision without considering the dog's features. Similarly, $\mathcal{T}_q^c$ is of the form "What if the tiger cat was not in the image?" Hence, a function that maximizes these terms, i.e., $\mathcal{M}_{\text{cf}}(\cdot)$, are counterfactuals from the "Observed Counterfactual Explanations" section. Therefore, any explanatory map that maximizes the probability term with $\mathcal{T}^c$ is an observed counterfactual explanation. Similarly, $\mathbb{P}(\mathcal{T}_p \mid P)$ and $\mathbb{P}(\mathcal{T}_q \mid Q)$ probability terms are maximized by functions that are observed correlation and contrastive explanations from the "Observed Correlation Explanations" and "Observed Contrastive" sections, respectively. The sum of these functions maximizes all probabilities for data x, and hence is a complete explanation $\mathcal{M}_c(x)$. Substituting these functions in (6), and using Definition 3.1, we have

$$\mathcal{M}_c(x) = \mathcal{M}_{\text{cu}}(x) + \mathcal{M}_{\text{ct}}(x) + \mathcal{M}_{\text{cf}}(x), \tag{7}$$

Hence, a complete explanation that accounts for all probabilities in a network simultaneously answers the three observed correlation, counterfactual, and contrastive questions. In other words, an explanatory technique that answers observed correlation, counterfactual, and contrastive questions is sufficiently complete. In this article, we advocate for creating complete explanations. We show that a combination of gradient-based methods from [12] and [13] can be used to achieve such an explanation in the "Gradient-Based Complete Explanations" section. In the "Significance of Observed Explanatory Paradigms and Complete Explanations" section, we elaborate on the significance of a complete explanation.

Gradient-based complete explanations

In Figure 3, the requisite change in manifolds for observed correlation, counterfactual, and contrastive explanations occurs because of the change in network parameters that spans the manifolds. From (3), the change in parameters can be characterized by gradients. In this section, we show that a gradient-based technique can provide a complete explanation. The authors in [12] propose Grad-CAM, an explanatory technique that provides a gradient-based localization map leading to a decision. Grad-CAM is also used to create counterfactual explanations by altering gradients. Grad-CAM is further extended as Contrast-CAM in [13] and answers contrastive questions. All these works use the same underlying explanatory mechanism, the difference is in the gradients used within each technique to access the requisite manifolds. Each of these techniques is elaborated on further in the next sections.

Grad-CAM

Grad-CAM [12] is used to visually justify the decision P made by a classification network by answering "Why P?" The activations from the last convolutional layer of a network are used to create these visualizations as they possess high-level semantic information while retaining spatial information. For any class $i, \forall i \in [1, N]$, logit y_i is backpropagated to feature map A_l, where l is the last convolutional layer. The gradients at every feature map are $\partial y_i / \partial A_l^k$ for a channel k. These gradients are global-average pooled to obtain importance scores α_k^{cu} of every feature map in the lth layer and kth channel. The individual maps A_l^k are multiplied by their importance scores α_k and averaged to obtain a heat map. The Grad-CAM map at layer l for class i is given by $\mathcal{M}^i = ReLU(\Sigma_{k=1}^{K} \alpha_k^{\text{cu}} A_l^k)$.

Counterfactual-CAM

The authors in [12] extend Grad-CAM by negating the gradients as $-\partial y_i / \partial A_l^k$, thereby decreasing the effect of the predicted class. The resulting importance score α_k^{cf} de-emphasizes the region used for classification. Effectively, these regions will be eliminated, and the highlighted regions are the ones that the network would have used to make its decision if the object used to make the decision were not present.

Table 1. Explanation map categorizations.

Paradigms	Methods	Definition			Technique Categorization						Evaluation		
		Indirect	Direct	Targeted	Implicit	Explicit	Black Box	White Box	Intervention	Nonintervention	Gradient Based	Nongradient Based	Direct Human	Indirect Application	Targeted Network
Correlation	Deconvolution [21]	✓	—	—	—	✓	—	✓	—	✓	—	✓	✓	—	—
—	Inverted Representations [22]	✓	—	—	—	✓	—	✓	—	✓	—	✓	✓	—	—
—	Guided-Backpropagation [18]	—	✓	—	✓	—	—	✓	—	✓	✓	—	—	✓	—
—	SmoothGrad [17]	—	✓	—	✓	—	—	✓	✓	—	✓	—	—	✓	—
—	LIME [39]	—	✓	—	—	✓	✓	—	✓	—	—	✓	✓	✓	—
—	CAM [24]	—	✓	—	—	✓	—	✓	—	✓	✓	—	—	✓	—
—	Graph-CNN [23]	✓	—	—	—	✓	—	✓	—	✓	—	✓	✓	✓	—
—	GradCAM [12]	—	—	✓	✓	—	—	✓	—	✓	✓	—	✓	✓	—
—	TCAV [40]	—	✓	—	—	✓	—	✓	—	✓	✓	—	✓	✓	—
—	GradCAM++ [16]	—	—	✓	✓	—	—	✓	—	✓	✓	—	✓	✓	—
—	RISE [35]	—	✓	—	✓	—	✓	—	✓	—	—	✓	—	✓	✓
—	Causal-CAM [15]	—	—	✓	✓	—	—	✓	—	✓	✓	—	✓	—	✓
Counterfactual	Counterfactual-CAM [12]	—	—	✓	✓	—	—	✓	—	✓	✓	—	✓	—	—
—	Goyal et al. [26]	—	—	✓	✓	—	—	✓	✓	—	—	✓	✓	✓	—
Contrastive	CEM [29]	—	—	✓	—	✓	—	✓	✓	—	—	✓	✓	✓	—
—	Contrast-CAM [13]	—	—	✓	✓	—	—	✓	—	✓	✓	—	✓	—	—
—	Contrastive reasoning [14]	—	—	✓	✓	—	—	✓	—	✓	✓	—	✓	—	✓

CNN: convolutional neural network.

Hence, it is of the "What if?" modality. The final counterfactual explanatory map at layer l is given by $\mathcal{M} = ReLU(\Sigma_{k=1}^{K} \alpha_k^{\text{cf}} A_l^k)$.

Contrast-CAM

In Grad-CAM, the importance score α_k^{cu} weighs the activations in a layer l based on A_l^k's contribution toward classification. The activations are projections on network parameters and hence have access to both correlation and contrastive information. The goal is to obtain contrast-importance score α_k^{ct}, which highlights the contrastive information in A_l^k. First, a loss between predicted class P and any contrast class Q is constructed as $J(P, Q)$. This loss function inherits the difference between classes P and Q. Next, $J(P, Q)$ is backpropagated within the Grad-CAM framework to obtain α_k^{ct} as global-average pooled contrastive features, i.e., $\alpha_k^c = \Sigma_u \Sigma_v \nabla_{W_l} J(P, Q)$, where u and v are the channel sizes at layer l. A Contrast-CAM map [13] for contrast class Q is given by $\mathcal{M}^i = ReLU(\Sigma_{k=1}^{K} \alpha_k^{\text{ct}} A_l^k)$.

Significance of observed explanatory paradigms and complete explanations

Results on noninterventionist data

In this section, we elaborate on the significance of all three paradigms of observed explanations. The results shown in Figures 1 and 2 are natural images when $f(\cdot)$ is a VGG-16 architecture. The observed correlation and counterfactual explanations are extracted using [12], and observed contrastive is extracted through Contrast-CAM [13]. In Figure 4, a fault in the seismic section of The Netherlands' F3 block [28] is analyzed. On such data, active interventions from a causal perspective are not possible. Instead, we rely on our observed paradigms. The data set [28] has four classes: faults, salt domes, chaotic regions, and horizons. An 18-layer residual network is trained, and the corresponding Grad-CAM [12], Counterfactual-CAM [12], and Contrast-CAM [13] explanations extracted from the trained model are shown. The observed correlation explanation "Why fault?" highlights the region where faults are recognizable. The observed contrastive explanation to the question, "Why fault, rather than salt dome?" performs the harder task of tracking the fault as the Grad-CAM explanation can also be confused to be highlighting the salt dome region. The counterfactual explanation answers "What if the fault did not exist?" and highlights the horizons class, indicating that horizons would be the classification. All three paradigms provide new insights to seismic interpreters, who are then qualified to assess the final decision. Note the complementary nature of the explanations. All

features are highlighted across the seismic section, thereby visually validating (7).

Using complete explanations to differentiate context features from observed correlation features

The authors in [15] use the completeness of observed gradient-based methods to provide a methodology to separate out context features $\mathcal{T}_c$ from observed correlation features $\mathcal{T}_p$. They analyze Grad-CAM and postulate that the Grad-CAM explanation "Why *P*?" is a combination of causal and context features, making it a correlation. They use a set-theoretic approach to provide relationships between the two sets of features and define context features in terms of contrastive and counterfactual features. This work can be considered an explanatory bootstrap, where multiple methods are tied together to produce a new explanation. Such a bootstrap is possible because each individual approach answers a specific question.

Using gradient features for applications

Posthoc explanations are passive in the sense that they justify decisions once they are made. However, observed contrastive and counterfactual explanations can actively be used for inference and training. The authors in [29] construct contrastive explanations to mine pertinent negatives to train networks. The authors in [14] use contrastive explanations and features to create robust networks. They show that use of contrastive features $\mathcal{T}_{p,q}$ in addition to observed correlation features $\mathcal{T}_p$ increases the robustness of a network to noise by 4% on the `CIFAR-10C` [30] data set. This robustness is also showcased on domain-shifted data, where accuracy increases by 5.2% on the `Office` [31] data set. The authors use these contrastive $\mathcal{T}_{p,q}$ features as a plug-in on top of existing $\mathcal{T}_p$ features, thereby heuristically validating the significance of complementary observed paradigms.

Analyzing explanatory techniques through observed paradigms

By defining explanations in terms of probabilities, we tie existing explanatory techniques to reasoning and complete them based on abstract questions that use learned features $\mathcal{T}$. This provides insights into existing explanatory techniques, along with a unique avenue for analyzing trends in the field of explainability. The comparisons of existing methods is performed in Table 1. We first provide a taxonomy of evaluation for the observed paradigms before analyzing the trends. We then showcase challenges in existing explainability techniques as analyzed through the lens of our question-based paradigms.

Evaluating explanations

The subjectivity of explainability creates a challenge for its evaluation. For a number of applications including recognition, the only available ground truths are object labels. Hence, explainability is not a task with ground-truth labels that can be easily evaluated. The authors in [4] provide a broad evaluation overview for AI techniques. We adapt this framework for neural network explainability and present it as an explanatory evaluation taxonomy. An overview is provided in Figure 5. Inherently, explanations are for humans. All evaluation criteria must be grounded in either direct or indirect interactions with humans. Consequently, there are three possible taxonomies that existing explanatory techniques use for evaluation. We first introduce and elaborate on these taxonomies. Specific explanatory paradigm techniques and their related evaluation criteria are provided in the "Categorizations, Trends, and Insights Into Explainability" section.

Direct human evaluation

In this taxonomy, explanations serve to directly convince humans of the decisions made by neural networks. As such, subjectively assessing the quality of explanations involves

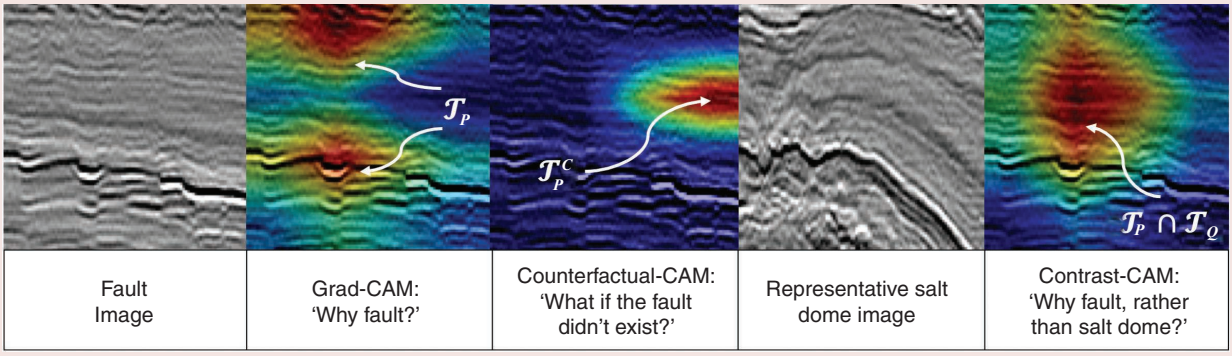


FIGURE 4. Observed correlation, counterfactual, representative contrast class and contrastive explanations, respectively on a fault image. The *'Why Fault?'* is answered by highlighting the discontinuity features $\mathcal{T}_P$ in the rocks that correlates to presence of a fault. The counterfactual explanation analyzes the decision after removing $\mathcal{T}_P^c$. The contrastive explanation highlights features that are an intersection of $\mathcal{T}_P$ and $\mathcal{T}_Q$. Here, *P* is the fault class and *Q* is the salt dome class.

humans in the loop. With the widespread applicability of neural networks, the most common evaluation of explainability is qualitative direct human evaluation. We show application results that $f(\cdot)$ was trained for, like recognition accuracy, along with selected explanations for images. This is true for all data domains, including natural images [21], medical images [15], and seismic images [5], among others. The early explainability techniques including [17], [18], and [21] all follow this approach. A quantitative approach to direct human evaluation is shown in [12]. The authors in [12] create explanation maps on 90 random images from the ImageNet data set and use Amazon Mechanical Turk to find and ask people to rate their explanation maps as compared to the explanations created by [18] and [24]. Note that the goal here is not to evaluate the efficacy of the explanatory map by itself, but rather compare between explanatory methods.

Although subjectively evaluating explanation maps is resource intensive, it provides new insights into data and applications. In the application of image quality assessment (IQA), the authors in [13] show that contrastive explanations provide additional context to the explanations. In Figure 6, P is the estimated quality score, and Q is a given contrast score. When $Q > P$, the contrastive explanations describe why a lower score is predicted. According to the obtained visualizations when $Q = 1$ and $Q = 0.75$, the estimated quality is primarily due to distortions within the foreground portions of the image as opposed to the background. A further analysis is provided in [13]. As direct human evaluations provide new qualitative information, Temel et al. constructed contrastive features from an existing IQA metric [32] and showed that quantitative results increase across multiple IQA data sets [33].

Indirect-application evaluation

In this taxonomy, tasks are designed that indirectly define explainability. Human evaluation is conducted on such tasks and because explainability is defined as a function of the task, explainability is evaluated simultaneously. The advantage of such an approach is twofold: 1) the existing framework for the application can be utilized and 2) explanations are not directly measured, which validates the original supposition that explanations are not a task by themselves but rather auxiliaries that support the reasoning behind tasks.

Evaluation tasks

In the application of recognition, the authors in [24] define *explanations* as localization maps of a given object within the image. They evaluate explanations as localization, fine-grained recognition, weakly supervised localization, pattern discovery, and visual question answering. The authors in [16] derive explanation $\mathcal{M}$ for an image x and mask x as $x' = x \times \mathcal{M}$. They pass the masked image through the network and check for accuracy. This is conducted across a data set of images for a number of explanatory techniques. The higher the accuracy, the better the localization capability of the explanation, and hence, the better the explanation. Note that here, the definition of *explanation* is the localization of objects. The authors in [34] use a visual saliency application as a proxy for explanation. The intuition is that the human gaze lingers on the most salient parts of a given image. These salient regions are explanations for decisions made on that image as it attracts the human gaze. The authors construct explanation maps using existing techniques of Grad-CAM [12] and guided backpropagation [18] and show that their own implicit saliency tracks the human gaze with a higher correlation. The other applications include weak segmentation [12], a pointing game [12], [35], and image captioning [35].

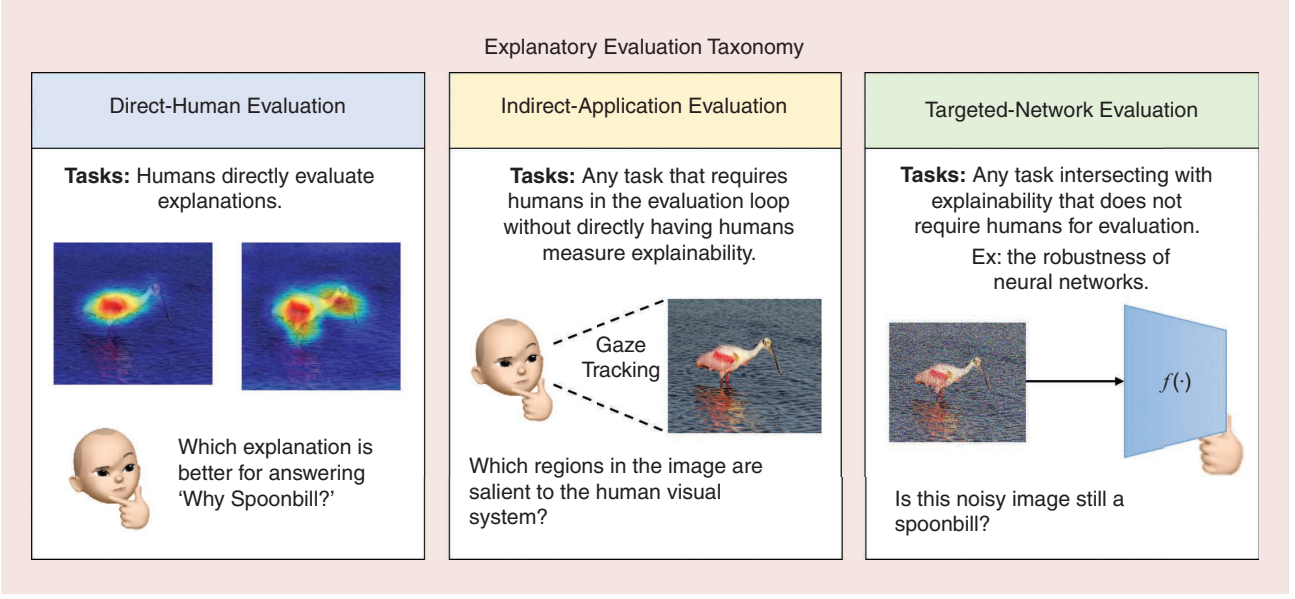


FIGURE 5. The three taxonomies of evaluation based on human involvement. The first is a direct evaluation of explanations by humans. The second defines explanations as proxy tasks that require human evaluation. The third uses existing tasks that have been shown to require interpretability. (Source: [42] (spoonbill).)

Evaluation metrics

The novelty in indirect-application evaluation is in aligning the definition of explanations with existing tasks. Once they are aligned, established metrics for each application are used for evaluation. For instance, classification accuracy is used to evaluate the masked images in [16]. Normalized scanpath saliency and a correlation coefficient are used in [34] for evaluating visual saliency.

Targeted-network evaluation

The authors in [35] question the nature of explainability as a tool for human justification and instead suggest that explanations provide an insider's view into a neural network's decision-making process. Such a process need not conform to that of a human's. Explainability is connected to other applications like robustness. The goal in this taxonomy is to create experiments that increase performance on auxiliary applications, thereby increasing their explainability.

Evaluation tasks

In [14], the application of robustness is tied to explainability. The authors show observed correlation and contrastive explanations for noisy images and how having a large corpus of explanations can be robust to distortions. Consequently, the gradient features used in explanations are used to train a simple multilayer perceptron that recognizes noisy images in a more robust fashion compared to a feedforward network. The same robustness is observed for domain-adaptation experiments. These gradients are used in other robustness applications including anomaly [36], novelty [37], and out-of-distribution detections [38]. The gradients used in this fashion across applications connected to explainability provide an indirect validation of their explanatory capacity. The authors in [26] propose using counterfactual explanations for the application of machine teaching, where humans learn about bird categorizations from neural networks. Note that even though humans are involved, the task for humans is not to evaluate explanations; rather, they learn from explanations, and the evaluation measures how well humans learn.

Evaluation metrics

The authors in [35] and [15] propose three new evaluation criteria—all for the application of recognition. These include probabilistic deletion and insertion [35], accuracy deletion and insertion [15], and transference of features [15]. All three metrics, however, are for observed correlation explanations. There are currently no specific evaluation metrics designed for observed counterfactual or contrastive explanations. The challenge is currently to define tasks for these two paradigms, and the metrics follow from the tasks, similar to an indirect human evaluation.

Correlation, counterfactual, and contrastive evaluations

A complete summary of existing techniques and their evaluation strategies are presented in Table 1. These summaries

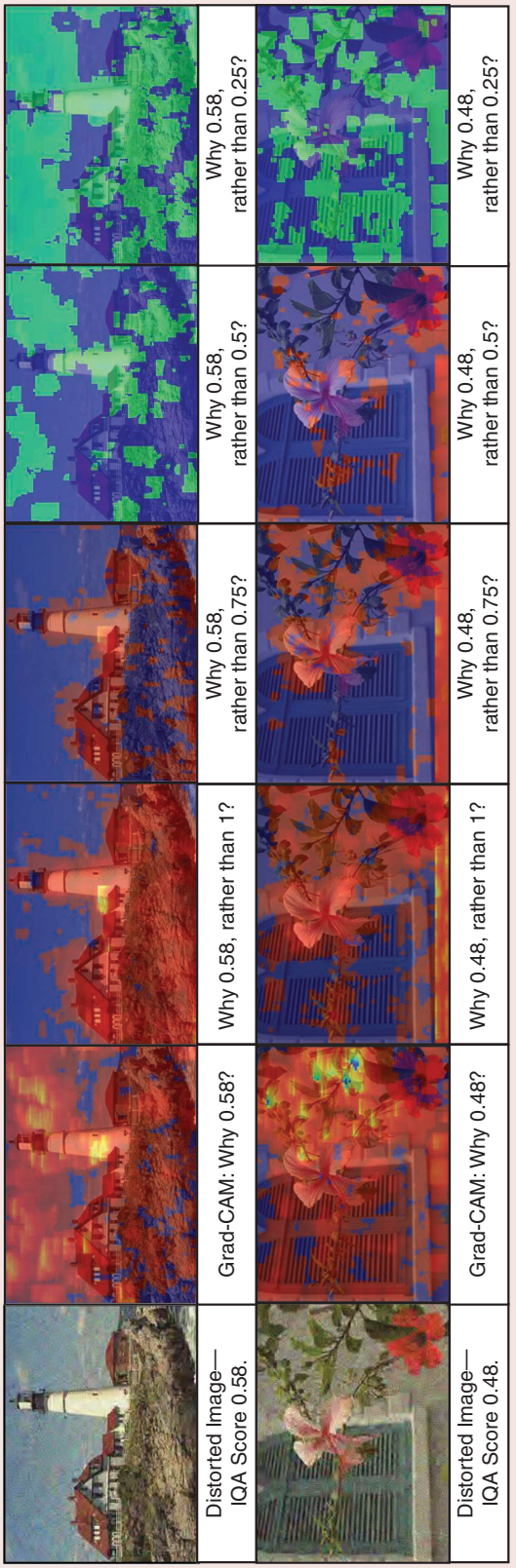


FIGURE 6. The Grad-CAM and contrastive explanations to questions shown below each image (best viewed in color). (Source: [13].)

provide insight into evaluation strategies followed in each of the observed correlation, counterfactual, and contrastive frameworks. The useful trends regarding evaluation, extracted from the table, are presented in the "Categorizations, Trends, and Insights Into Explainability" section.

Categorizations, trends, and insights into explainability

The explanatory paradigms are natural extensions of exiting works. Although categorization based on abstract explanations seem intuitive, existing techniques derive their novelty from other means. Specifically, existing explanatory categorizations are based on design choices within the explanatory techniques themselves. These design choices include a need for $f(\cdot)$'s parameters during explanations and an architectural design change in $f(\cdot)$ for deriving explanations, among others. We first expand on these categorizations before summarizing existing methods in Table 1.

Implicit versus explicit explanations

A number of techniques require an architectural change to the network or a separate network to generate explanations [21], [24], [29], [39], [40]. We term them *explicit explanations*. In some cases, a change in network architecture can potentially alter underlying results [12]. Implicit explanations follow the original definition of explanations, which is *why explanations justify inference*.

Black- versus white-box explanations

The techniques including [12], [13], [21], and [24] all require access to the model, its parameters, and/or network states to provide explanations. These are termed *white-box techniques*. The other methods that exist treat neural networks as black boxes that make decisions [35]. Both RISE [35] and LIME [39] require model outputs, not parameters.

Interventionist versus noninterventionist explanations

When an explanatory technique requires changes in data, we term them *interventionist* ones. SmoothGrad [17] requires noise added to the images. Counterfactual visual explanations [26] require parts of the input image substituted by another. RISE [35] requires generation of a large quantity of masked input images. Noninterventionist techniques do not require interventions in input data.

Gradient- versus nongradient-based explanations

A number of methods use gradients described in the "Neural Network Preliminaries" section as features. Some approaches, like SmoothGrad [17] and guided backpropagation [18], directly use gradients as explanation maps. The others, including Grad-CAM [12] and Contrast-CAM [13], use them as features within their frameworks.

All the existing methods are presented in Table 1. They are chronologically ordered within each paradigm. The rows ascertain memberships to paradigms, while there are three categorizations within columns. The first is the definitions of explanations from the "Background" section, the second is the technique-based categorization presented earlier in the section, and finally, the evaluation categorization from the "Evaluating Explanations" section. Note that in Table 1, a targeted definition is added to indirect and direct explanations. This includes all newer techniques related to the reasoning paradigms. The table provides key insights into the evolution of explainability approaches and potential research directions. These include the following:

- Correlation methods are studied more than the other explanatory paradigms. The indirect and direct definitions of explainability feeds observed correlations and has been well studied over time.
- Indirect and direct explanations have given way to targeted ones. *Targeted explanations* refer to those justifying contextually relevant questions.
- Of the 10 techniques that use direct human evaluation, only three, [12], [16], and [26], do so by conducting random trials on humans. The others are qualitative works whose images are potentially chosen by the researchers to showcase their methods. More rigor is required in direct human-evaluation strategies.
- In recent times, the difficulty in evaluation has encouraged authors to assess their works using multiple strategies. Since 2017, all but one work [13] use more than one taxonomy to evaluate explanatory techniques.
- There is no consensus among researchers regarding the targeted-network task that evaluates explainability. The primary task using explanation-masked images [16], as well as robustness tasks [14], are used in a majority of these evaluations.
- Although technique-based categorization provides some novelty to the methods themselves, they are ineffective in predicting trends in explainability over time.

Challenges in explainability and future research directions

Disparity between human and machine explanations

Consider the case in Figure 7(a) when a network, trained on the CURE-TSR [41] traffic sign data set, is analyzed constructively. Given a no-right-turn sign and when asked a contrastive question of the form "Why no left turn, rather than Stop?", the network highlights the bottom-left corner of the image. The letters that spell "Stop" not being in the sign is the intuitive human response to the aforementioned question. This disparity between human intuition and machine explanation is explained based on the data set. Among the 14 traffic signs in CURE-TSR, the stop sign is the only class that has a hexagonal shape. Hence, the network has learned to check for a straight side in the bottom left. The absence of this side in x indicates to the network that x is not a stop sign. Hence, the observed contrastive explanation clearly illustrates the disparity between the notion of classes between humans and machines, while observed

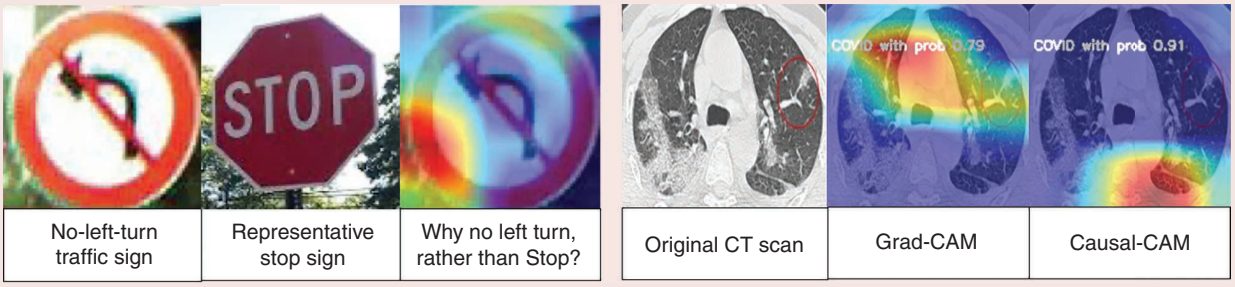


FIGURE 7. The contrastive explanation in (a) shows the disparity between human and machine explanations. Both methods in (b) fail to highlight the red circle. CT: computerized tomography. [Sources: [41] (traffic signs), [44] (CT scans).]

correlations could not. This necessitates consideration of complete explanations in the absence of interventions. It is not very clear whether the two explanations must align in all applications. However, for safety-critical applications in medical fields, it is required that these anomalies be analyzed and understood.

Failure of existing objective evaluation metrics

In Figure 7(b), we visualize Grad-CAM [12] and Causal-CAM [15] maps. The original scan is from a COVID-19-positive patient. Both the Grad-CAM and Causal-CAM explanations fail to highlight the circled red region that depicts COVID-19. Feeding the masked image back into $f(\cdot)$ for an indirect-application evaluation, the network classifies both correctly but with a higher confidence for Causal-CAM. Hence, in both the probabilistic and accuracy-deletion and insertion metrics, Causal-CAM performs better than Grad-CAM. This essentially shows that the network has learned to classify correctly but with the wrong features. Such a feature discrepancy is not reflected in the objective metrics of evaluation. In real-world biomedical applications like those in COVID-19 detection, it is imperative to identify and make decisions that are interpretable. Further study is required in designing better objective metrics and networks whose features are more human interpretable.

Conclusions

The key takeaways from this article include 1) realizing explanations as reasoning paradigms, 2) the definition of complete explanations in neural networks based on reasoning, and 3) utility of gradient-based explanatory paradigm's replicability across multifarious applications.

Explanations as reasoning paradigms

By having explanations answer targeted questions, the goal of explainability shifts from justifying decisions to reviewing and participating in making decisions. This is in line with an abductive reasoning scheme that humans often use in practice.

Complete explanations

Existing techniques treat each correlation, counterfactual, and contrastive paradigms as separate entities. However, each paradigm answers complementary questions. A complete explanation is one that can answer any given question. In practice, the combination of Grad-CAM [12], Counterfactual-CAM [12], and Contrast-CAM [13] achieve this. Similar research into methodologies that unify these paradigms and provide complete explanations is required and pertinent.

Applicability of gradients as features

From Table 1, it is clear that gradients are used as features in explanations across multiple categorizations and paradigms. The same gradients are also used as features for recognition robustness [14], anomaly [36], novelty [37] and out-of-distribution detection robustness [38], IQA [33], and visual saliency detection [34]. The adoption of a network's gradients as features across applications is a sign of its multifarious utility.

Acknowledgment

Ghassan AlRegib is the corresponding author.

Authors

Ghassan AlRegib (alregib@gatech.edu) received his Ph.D. in electrical engineering from the Georgia Institute of Technology (Georgia Tech), Atlanta, Georgia, 30332, USA. He is currently the John and Marilu McCarty Chair Professor in the School of Electrical and Computer Engineering at the Georgia Institute of Technology, Atlanta, Georgia, 30332, USA. His research group, the Omni Lab for Intelligent Visual Engineering and Science (OLIVES), works on research projects related to explainable machine learning, robustness in intelligent systems, interpretation of subsurface volumes, and expanding health-care access and quality. AlRegib has participated in several service activities within IEEE and served on the editorial boards of several journal publications. He served as the technical program cochair for the 2020 IEEE International Conference on Image Processing and the 2014 IEEE Global Conference on Signal and Information Processing. AlRegib is a Fellow of IEEE.

Mohit Prabhushankar (mohit.p@gatech.edu) received his Ph.D. degree in electrical engineering from the Georgia Institute of Technology (Georgia Tech), Atlanta, Georgia, 30332, USA, in 2021. He is currently a postdoctoral researcher in the Omni Lab for Intelligent Visual Engineering and Science, working in the fields of image processing and machine learning. He has been a teaching fellow at Georgia Tech since 2020. He is the recipient of the Best Paper Award at the 2019 IEEE International Conference on Image Processing (ICIP 2019) and the Top Viewed Special Session Paper Award at ICIP 2020. Prabhushankar is a Member of IEEE.

References

[1] A. A. Freitas, "Comprehensible classification models: A position paper," *ACM SIGKDD Explorations Newslett.*, vol. 15, no. 1, pp. 1–10, 2014, doi: 10.1145/2594473.2594475.

[2] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomed. Eng.*, vol. 2, no. 3, pp. 158–164, 2018, doi: 10.1038/s41551-018-0195-0.

[3] S. Carton, J. Helsby, K. Joseph, A. Mahmud, Y. Park, J. Walsh, C. Cody, C. E. Patterson *et al.*, "Identifying police officers at risk of adverse events," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 67–76, doi: 10.1145/2939672.2939698.

[4] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[5] M. A. Shafiq, M. Prabhushankar, H. Di, and G. AlRegib, "Towards understanding common features between natural and seismic images," in *SEG Tech. Program Expanded Abstracts 2018*. Society of Exploration Geophysicists, 2018, pp. 2076–2080. [Online]. Available: https://library.seg.org/doi/abs/10.1190/segam2018-2996501.1

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[7] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam *et al.*, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[9] P. Kitcher and W. C. Salmon, *Scientific Explanation*. Minneapolis, MN, USA: Univ. of Minnesota Press, 1962, vol. 13.

[10] J. Pearl, "Causal inference in statistics: An overview," *Statist. Surv.*, vol. 3, pp. 96–146, Sep. 2009, doi: 10.1214/09-SS057.

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986, doi: 10.1038/323533a0.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[13] M. Prabhushankar, G. Kwon, D. Temel, and G. AlRegib, "Contrastive explanations in neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2020, pp. 3289–3293, doi: 10.1109/ICIP40778.2020.9190927.

[14] M. Prabhushankar and G. AlRegib, "Contrastive reasoning in neural networks," 2021, *arXiv:*2103.12329.

[15] M. Prabhushankar, and G. AlRegib, "Extracting causal visual features for limited label classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2021, pp. 3697–3701.

[16] A Chattopadhay, A Sarkar, P Howlader, and V N Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[17] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[18] J. Tobias Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*.

[19] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, 2021, doi: 10.1109/JPROC.2021.3058954.

[20] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Scholkopf, and L. Bottou, "Discovering causal signals in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6979–6987, doi: 10.1109/CVPR.2017.14.

[21] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[22] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5188–5196, doi: 10.1109/CVPR.2015.7299155.

[23] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, and S.-C. Zhu, "Interpreting CNN knowledge via an explanatory graph," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4454–4463.

[24] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929, doi: 10.1109/CVPR.2016.319.

[25] K. Chalupka, P. Perona, and F. Eberhardt, "Visual causal feature learning," 2014, *arXiv:1412.2309*.

[26] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," 2019, *arXiv:1904.07451*.

[27] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, 2018, pp. 80–89, doi: 10.1109/DSAA.2018.00018.

[28] Y. Alaudah, M. Alfarraj, and G. AlRegib, "Structure label prediction using similarity-based retrieval and weakly supervised label mapping structure label prediction," *Geophysics*, vol. 84, no. 1, pp. V67–V79, 2019, doi: 10.1190/geo2018-0028.1.

[29] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," 2018, *arXiv:1802.07623*.

[30] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019, *arXiv:1903.12261*.

[31] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 213–226, doi: 10.1007/978-3-642-15561-1_16.

[32] D. Temel, M. Prabhushankar, and G. AlRegib, "Unique: Unsupervised image quality estimation," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1414–1418, 2016, doi: 10.1109/LSP.2016.2601119.

[33] G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Distorted representation space characterization through backpropagated gradients," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2019, pp. 2651–2655, doi: 10.1109/ICIP.2019.8803228.

[34] Y. Sun, M. Prabhushankar, and G. AlRegib, "Implicit saliency in deep neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2020, pp. 2915–2919.

[35] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," 2018, *arXiv:1806.07421*.

[36] G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated gradient representations for anomaly detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 206–226.

[37] G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Novelty detection through model-based characterization of neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2020, pp. 3179–3183, doi: 10.1109/ICIP40778.2020.9190706.

[38] J. Lee and G. AlRegib, "Gradients as a measure of uncertainty in neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2020, pp. 2416–2420, doi: 10.1109/ICIP40778.2020.9190679.

[39] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[40] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2668–2677.

[41] D. Temel, G. Kwon, M. Prabhushankar, and G. AlRegib, "CURE-TSR: Challenging unreal and real environments for traffic sign recognition," 2017, *arXiv:1712.02463*.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[43] N. Ponomarenko *et al.*," Image database TID2013: Peculiarities, results and perspectives, *Signal Processing: Image Commun.*, vol. 30, pp. 57–77, 2015. [Online]. Available: https://doi.org/10.1016/j.image.2014.10.009

[44] J. Zhao, Y. Zhang, X. He, and P. Xie, "Covid-ct-dataset: a ct scan dataset about covid-19," vol. 490, 2003, *arXiv:*2003.13865.

SP