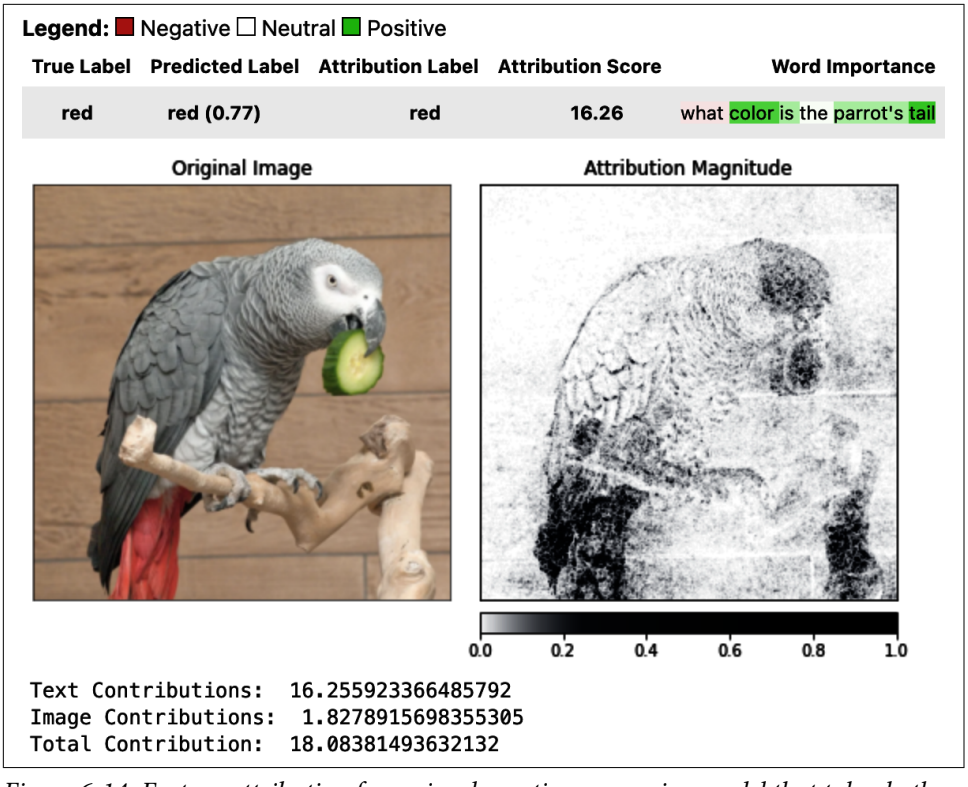

Figure 6-14. Feature attribution for a visual question answering model that takes both an image and a question (text) as inputs. The highlights show which words and which pixels contributed the most to the predicted label. Feature attributions are relative, and aggregation can be done to get the modality level attribution that yields a contribution of 16.26 from the question and 1.83 from the image. (Print readers can see the color image at https://oreil.ly/xai-6-14.)

Evaluation of Explainability Techniques

For most part, this book has focused on well-established explainability techniques that have been widely used across different applications. At this stage, you're probably wondering how these techniques compare with each other or if one technique is better than the others somehow. Unfortunately, there is no free lunch and different techniques may do better or worse depending on your dataset, your use case, and how you plan to use the resulting explanations. Just as there is no one-size-fits-all machine learning model, there is no single explainability technique that surpasses all the rest. Instead, we encourage you to view and collect these techniques as tools in a well-stocked toolkit that exists to help you analyze your model and can be utilized through your entire ML workflow (see Chapter 8).

Although you may have a bunch of explainability techniques at your disposal, the question still remains as to how you can evaluate them and which one is best to use for your use case. Evaluating a predictive machine learning model is much more straightforward; there are well-known metrics you can use like accuracy, precision/recall, mean-square error, intersection over union, and so on. However, there is a lack of consensus for how to evaluate the quality of explanations. Often, researchers have relied on showing a handful of examples to convince others of the usefulness of their proposed explanation technique, but over time more attention has been devoted to developing systematic evaluation methods. In this section, we will go over a few different approaches to evaluating explainability techniques. Even though this discussion is not exhaustive, we aim to provide you with a starting point and to create awareness for the topic.

A Theoretical Approach

With the lack of clear evaluation metrics for assessing explainability techniques, one approach is to take a "first principles" perspective, meaning with no preconceived assumptions, and define a collection of axioms that any practitioner would expect an explainability technique to have.

Axioms of Mathematics

In mathematics, axioms are statements that cannot be proven true or false. They are accepted to be self-evidently true and serve as the starting point from which the rest of the abstract theory can be developed. For example, the axiom of equality states that a number is always equal to itself. Or perhaps, in more layman's terms, "It is what it is." This seems too obvious to not be true, and it's what any reasonable person would expect. However, this statement can't formally be proven true or false, so it is accepted as true. Axioms like this form the foundation of mathematical theory, from which the proofs of all other theorems, propositions, lemmas, and conjectures follow.

An axiomatic approach defines a collection of well-understood and accepted properties (i.e., axioms) that any explainability technique *should* have. This provides an intuitive benchmark with which to judge new or existing techniques. In their 2017 paper[19] introducing the technique of Integrated Gradients, Sundararajan et al. also introduce a collection of fundamental axioms for qualifying feature attributions. We'll briefly discuss those axioms here because they provide a nice sanity check for what

19 Mukund Sundararajan et al., "Axiomatic Attribution for Deep Networks," International Conference on Machine Learning, PMLR, 2017.

we expect an explainability method should possess and, when possible, we'll compare these axioms against some of the techniques we've seen so far in the book. However, keep in mind that these axioms are just one way of framing evaluation, and just because one method doesn't satisfy one of the axioms doesn't mean it should be abandoned completely. It's very likely that it could still be a useful technique for you and for your use case.

To frame the following axioms, we'll borrow the notation from the 2017 paper. Let F represent the model's (e.g., a deep neural network) prediction for a certain class label and the $x = (x_1,x_2,\cdots,x_n)$ represent an input tensor to F. We measure attributions in the context of a baseline, so let x' denote the baseline. For example, suppose we have some input image, as in Figure 6-15, and our model predicts the class label "sulfur-crested cockatoo" with a confidence of 0.886.

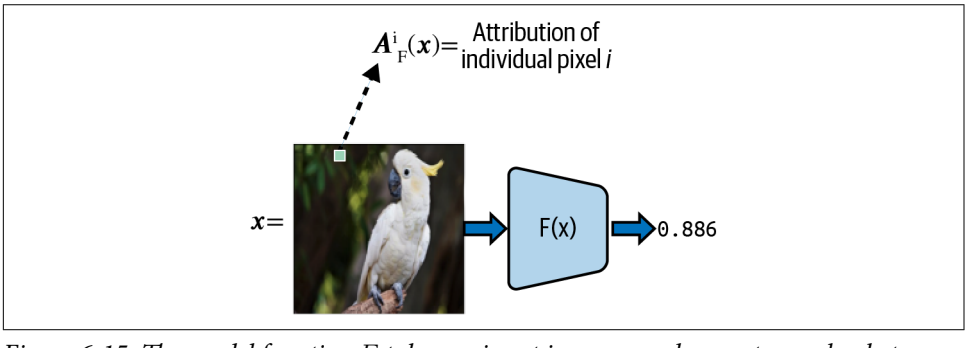

Figure 6-15. The model function F takes an input image x and maps to a value between 0 and 1, which represents the model's class prediction for that example.

The model function F maps the input image x to a probability score in (0,1). If we take the baseline to be a black image, then $F(x') = 0$. The features of the model are the pixels, the individual x_i's of the image, so we take $A^i_F(x)$ to represent the feature attribution of pixel x_i of the model F at the input x with respect to the baseline x'.

Axiom of completeness

The axiom of completeness is perhaps the most straightforward. It states that for any input to the model, the total attribution must be equal to the sum of all of the feature attributions of the input. Mathematically, this means:

$$F(x) - F(x') = \sum_i A^i_F(x)$$

You can think of completeness as a sanity check that the attribution method is comprehensive in assigning attributions to features. That is to say, the attribution of a given input is "completely" accounted for across all model features. It seems like

a reasonable request, and many of the explainability methods we've discussed satisfy this criteria, such as Integrated Gradients, Shapley values, and Layer-Wise Relevance Propagation. However perturbation techniques, such as LIME, do not.

Axiom of sensitivity

As the name suggests, the axiom of sensitivity examines how sensitive or stable an explainability method should be to changing feature values. It states that if the baseline and input differ only by one feature, but the prediction is the same, then that feature should have zero attribution. And conversely, if the input and baseline differ by only one feature, and the prediction is different, then that feature must have nonzero attribution. For example, in Figure 6-16, the two images differ by a single pixel value, but the model prediction doesn't change. The (in)sensitivity axiom says that in this case the attribution for that feature (i.e., pixel) should be zero.

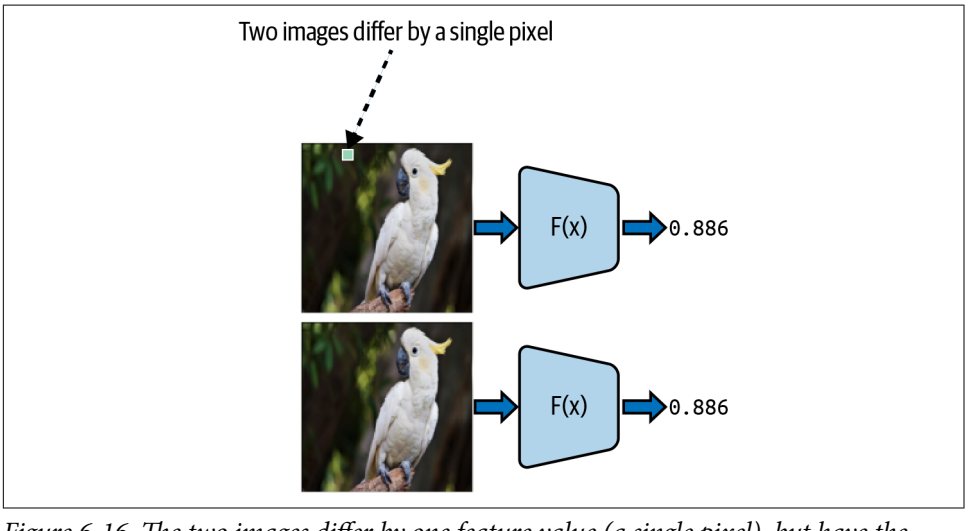

Figure 6-16. The two images differ by one feature value (a single pixel), but have the same prediction. The feature attribution for this pixel should be zero.

Mathematically, this axiom can be formulated as follows: if an input example x has only one nonzero feature and $F(x) \neq 0$ then the attribution for that feature is zero. Stated with respect to insensitivity, this axiom states that if the model function $F(x)$ does not depend on the value of a feature, as in Figure 6-16, then the attribution of that feature should be zero.

You can show that completeness implies sensitivity, so any technique that satisfies the axiom of completeness also satisfies the sensitivity axiom. Although perturbations methods of explainability do not satisfy completeness, they do satisfy sensitivity. However, methods like DeConvNets and Guided Backprop, which we discuss in Chapter 4, violate this sensitivity axiom.

Axiom of implementation invariance

The next axiom is related to the dependence on the machine learning model itself. The axiom of implementation invariance states that if two different models are functionally equivalent, meaning they compute the exact same function $F(x)$, then the attributions for all the features should be the same. In short, if the function F doesn't change between two models, then their attributions shouldn't change either. This axiom seems pretty straightforward to expect of an explainability technique. After all, why should the attributions change just because the implementation changes? So long as the final prediction is the same, that's all that should matter, right? Well, there are indeed methods that don't satisfy this criteria, such as DeepLIFT and Layer-Wise Relevance Propagation.

Axiom of linearity

The axiom of linearity says that if you can express your machine learning model as a linear combination of two other model functions, then the attributions should also be expressed in the same way. For example, if the model function F can be written as the sum of two model functions then you should expect that the attributions sum as well. Integrated Gradients and other path-based methods satisfy this axiom.

Axiom of symmetry-preserving

The last axiom is the axiom of symmetry-preserving. We say that two features are symmetric if they can be interchanged and the model function F prediction doesn't change. The axiom states that if two features are symmetric, meaning they are interchangeable without changing the value of F, then their attributions should be the same as well. This is somewhat related to the sensitivity axiom, but they're slightly different. The axiom of sensitivity says that if you change one feature value and the total attribution of the input doesn't change, then that feature must have zero attribution. Symmetry-preserving instead pertains to the model function itself. It means that there are two input variables that are symmetric. So, for any image, swapping these two pixel values wouldn't change the predicted value of F. In this case, the attributions should be the same as well.

This axiomatic approach to evaluating explainability techniques provides a reasonable sanity check and a good starting point. However, just because a certain technique fails one of these axioms doesn't mean you should discount it completely. Depending on your use case, that technique could still be beneficial to you. They simply provide one lens with which to understand the potential drawbacks or caveats of a certain method. Since these axioms were introduced, there have been a number of other studies aimed at providing a more rigorous framework for evaluating XAI techniques. Next, we'll discuss some of these empirical approaches.

Empirical Approaches

In the previous section, we looked at some of the desirable properties of an explanation technique and stated them as axioms establishing a theoretical framework. Trying to prove that your explanation method satisfies these axioms can get mathematically cumbersome and might be overkill if the aim is to quickly establish whether the method is good enough for your use case. In this section, we look at empirical approaches that can be applied to most techniques out of the box and are founded on intuition about how a technique should behave in simple scenarios.

Basic sanity checks

Basic sanity checks are the first line of investigation for the subject of evaluation. A sanity check is a quick way to weed out poor methods. A good method must necessarily pass such a check, but passing the check is not sufficient proof of correctness or good performance, since the checks are not intended to be exhaustive. With this in mind, one of the simplest questions to ask is whether an explainability technique truly does reveal something about the underlying model behavior. If the explanations don't change much with different models, the quality of the explainability technique appears rightfully suspicious. Similarly, if the data labels were changed, breaking the relationship between the model inputs and their outputs, we should expect explanations to change as well.

A sanity check–based approach[20] has been used to evaluate saliency maps for commonly used explainability techniques, including many of the methods we've discussed in this book like Guided Backprop, Guided Grad-CAM, Grad-CAM, Integrated Gradients, and Gradient x Input. To test the sensitivity of these methods to the model weights, Adebayo et al. compare the output of each saliency method when applied to a trained model against the same saliency output for the model with randomized weights. To test for sensitivity to data, a data randomization test compares the outputs of these saliency methods for a model trained on the true dataset and for a model trained on a copy of the dataset where the labels have been randomly shuffled. As with the model parameter randomization test, you would expect that the saliency outputs would change dramatically when the labels are randomized.

Surprisingly, many of these saliency methods failed these basic sanity checks. For example, the saliency maps produced by Guided Backprop and Guided Grad-CAM were found to be not sensitive to these randomization tests. A follow-up paper[21] discovered some gaps in the evaluation scheme of Adebayo et al., suggesting caveats on

20 Julius Adebayo et al., "Sanity Checks for Saliency Maps," *Advances in Neural Information Processing Systems* 31, 2018.

21 Mukund Sundararajan and Ankur Taly, "A Note About: Local Explanation Methods for Deep Neural Networks Lack Sensitivity to Parameter Values," arXiv, 2018.

how sensitivity should be measured and how visualization of the results can induce bias. These sanity checks, when done properly, can still be of value and provide one means of evaluation for an explainability technique. As a practitioner, you should be cautious when interpreting the results of both the explainability techniques and the evaluation schemes.

Faithfulness check

The next and a more challenging way to evaluate explainability techniques is to consider faithfulness of explanations to the model; that is, whether high attribution (to features or training data points) actually implies importance. In other words, an explainability technique can be considered to be revealing a model's strong reliance on a feature or training data point, if the model's performance suffers strongly when the high attribution inputs (features or training data points) are removed. However, if a technique assigns a high attribution to an input (feature or a training data point), and removing or altering that input doesn't change the model's performance significantly, the technique's attribution is less reliable.

In the paper that introduces XRAI,[22] a masking-based evaluation scheme is defined that captures this idea: starting with a blurry image, most- to least-salient pixels (based on the attribution) are sequentially introduced and model performance is evaluated. For a good explainability technique, we expect sharp improvement in performance at the start since the high saliency pixels are indeed the ones that would be the most useful to the model.

As a concrete example, let's consider a 5×5 image of a cat for which the model predicts "cat" with confidence 0.9. This means there are 25 pixels that can be ranked by attribution. Starting from a blurry image, we add pixels one by one based on their attribution values and notice how the prediction changes. Let's assume that the blurry image has a confidence score of 0 for the image being a cat. For a faithful technique, we would expect a big jump in confidence (say, from 0 to 0.3) when the highest attribution pixel is added, a somewhat smaller jump (e.g., from 0.3 to 0.5) for the next one, an even smaller jump (0.5 to 0.6) for the one after, and so on. This would be an insertion-based check. We can also do a deletion-based check by removing high attribution pixels one by one, and for a good technique, we should see the prediction fall sharply at the start. If the technique was poor, we should see erratic jumps with both the insertion or the deletion checks.

22 Andrei Kapishnikov et al., "XRAI: Better Attributions Through Regions," Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

Synthetic datasets

Another approach to evaluation is through synthetic datasets where the ML practitioner knows what the salient inputs are a priori. An explainability technique that can demonstrate it picks out these salient inputs is then more trustworthy than the one that picks out nonsalient inputs. For feature-based explanations, Yang, Mengjiao, and Kim provide a case study[23] on how synthetic data can be used for evaluating explainability techniques. By taking an image patch from a specific class (say, dog) and pasting it in different backgrounds, this approach ensures that the background is never quite salient since an image can be a dog anywhere (dog on a beach, dog in a gym, dog in a ring, etc.) and a good explainability technique shouldn't attribute much relevance to the background. For example-based explanations, similar augmentations can be made by adding random data points to the training set and measuring how attribution gets assigned to these nonsalient data points.

Application specific

Using basic sanity checks, faithfulness checks and checks built using synthetic datasets can help you evaluate your explanation technique in many scenarios. However, as you might have noticed, most of these checks are built as safeguards against obvious mistakes and don't guarantee success for a technique that has passed them. Another prevalent theme of explainability evaluation is centered around real use cases and becomes most relevant for sensitive applications. If explanations are being used to facilitate cancer diagnosis, the evaluation should consider metrics like time saved, mistakes avoided, new mistakes made, etc. A technique can pass all the previous checks, but if it doesn't save doctors any time while diagnosing, it might not be a great fit for the application. Similarly, if the explanations are being used to convey understanding of a model's inner workings to an everyday user, the evaluation should consider how well the explanations align with the user's intuition (this is discussed in more depth in Chapter 7).

Evaluation of explainability techniques is an active and hotly debated topic with a consistent flux of new ideas. As these explainability techniques continue finding their way into critical or regulated applications, the question of how much trust can be placed in any technique will continue to be pertinent. Unlike areas in machine learning where the notion of ground truth is well-defined leading to clear metrics, quantifying the quality of evaluation techniques is further complicated by the fact that for real-world datasets, the ground truth is often ambiguous—is that an image of a bee because of the striped pattern or because of the presence of specific flowers? Both are valid explanations even if one corresponds more closely to human intuition. Until we have well-established evaluation methodologies, it is worthwhile to spend

23 Yang, Mengjiao, and Been Kim, "Benchmarking Attribution Methods with Relative Feature Importance," arXiv, 2019.

some time ensuring alignment between the technique's intended use case and the application at hand.

Summary

In this chapter, we went over some of the emerging topics in the field of explainability. Different ways of approaching the question of explainability continue to receive attention from both researchers and practitioners. With that in mind, we first discussed how attribution or saliency can be assigned to inputs other than just the data features. Specifically, we looked at attributing to training data via a notion of similarity and influence functions. We also revisited attribution to concepts that can be defined via an auxiliary input.

Next we looked at a class of explainability methods that intervene during the modeling process as opposed to the post hoc techniques that are the main area of focus for the book. You saw how constraints like linearity and monotonicity can lead to more transparent and trustworthy models. We also went over how complex models can be distilled to simpler approximate models, often by using richer data generated using the complex model. Such models are easier to understand and have been successfully used for applications like ensuring fairness.

To reinforce the point that several of the techniques discussed in this book are model and modality agnostic, we looked at a couple of examples of applying Shapley values and integrated gradient techniques to different modalities, with the former being mostly applicable out of the box for time-series and the latter applicable to mixed (image and text) modality with low-overhead pre- and postprocessing.

Lastly, we looked at the fairly nascent field of systematic evaluation of explainability techniques. Historically, hand-picked examples of a technique's good performance were taken to be sufficient evidence for the quality of a technique. However, lately, there has been a larger push toward using well-defined, independent evaluation frameworks, especially as explainability continues to be applied to sensitive domains. Ranging from basic sanity checks to faithfulness studies to methods employing synthetic data, these frameworks aim to ensure alignment between the user's expectations and the technique's performance. The goals and methodologies for evaluation are not clearly understood, and extra care is required when using XAI for critical applications. Since a significant amount of XAI usage is by decision-makers, research is also experiencing a surge in field studies involving interactions between the explainability tools and human users. We continue to explore this aspect in the next chapter.