

# Exploring Principled Visualizations for Deep Network Attributions

Mukund Sundararajan  
Google Inc  
Mountain View, California  
mukunds@google.com

Shawn Xu  
Verily Life Sciences  
South San Francisco, California  
jinhuaxu@verily.com

Ankur Taly  
Google Inc  
Mountain View, California  
ataly@google.com

Rory Sayres  
Google Inc  
Mountain View, California  
sayres@google.com

Amir Najmi  
Google Inc  
Mountain View, California  
amir@google.com

## ABSTRACT

Attributions (cf. [1]) are increasingly used to explain the predictions of deep neural networks for various vision tasks. Attributions assign feature importances to the underlying pixels, but what the human consumes are visualizations of these attributions. We find that naive visualizations may not reflect the attributions faithfully and may sometime mislead the human decision-maker. We identify four guiding principles—Graphical Integrity, Layer Separation, Morphological Clarity, and Coverage—for effective visualizations; the first three requirements are standard in the visualization and the computer-vision literatures. We discuss fixes to naive visualization to satisfy these principles, and evaluate our fixes via a user study. Overall, we hope that this leads to more foolproof visualizations for mission-critical tasks like diagnosis based on medical imaging.

## CCS CONCEPTS

• **Human-centered computing** → **Heat maps; Visualization theory, concepts and paradigms**; • **Computing methodologies** → *Neural networks*.

## KEYWORDS

Deep Learning Interpretability; Prediction Explanation; Saliency

### ACM Reference Format:

Mukund Sundararajan, Shawn Xu, Ankur Taly, Rory Sayres, and Amir Najmi. 2019. Exploring Principled Visualizations for Deep Network Attributions. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019*, 11 pages.

## 1 INTRODUCTION

### 1.1 Visualization of Attributions

Deep neural networks are now commonly used for computer vision tasks. Such networks are used to detect objects in images (cf. [22]), and to perform diagnosis from medical imaging data. One approach to explaining the prediction of a deep network is to attribute the prediction score back to the base features (pixels). Several attribution methods have been proposed in the literature

(cf. [1, 2, 6, 14, 15, 17, 20], where they are also called local explanation methods or deep network saliency maps). The resulting attributions are meant to guide model or training data improvements, or assist a human decision-maker (such as a doctor using a deep learning model as assistance in the diagnosis of diabetic retinopathy [3, 12]).

Attribution methods fall into two broad categories. Some methods assign influence proportional to the gradient of the prediction score with respect to the input image (or modifications of the input image) (cf. [14, 16, 18]). Other methods propagate or redistribute the prediction score, layer by layer of the deep network, from the output of the network back to its input (cf. [2, 14, 17]). In all cases, the methods eventually assign each pixel a score proportional to its importance. This score could be positive or negative depending on the polarity of the influence of the pixel on the score. All the attribution methods are constructed and justified in principled ways (cf. the axiomatic approaches underlying [20] and [6], or the discussion about the Layerwise Conservation Principle from [11]). The justifications and the construction principles differ across methods because there seems to be no such thing as an universally optimal explanation, but nevertheless each method is relatively principled.

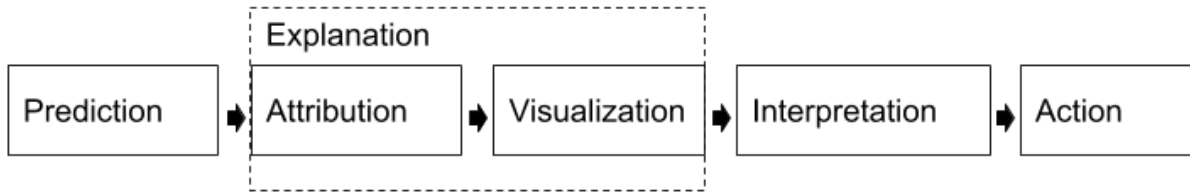
A key feature of these attribution based explanations is that they are expressly for human consumption. A common way to communicate these attributions to a human is by displaying the attributions themselves as images.<sup>1</sup> Figure 1 describes the context surrounding the visualizations.

### 1.2 Faithfulness of Visualizations

Because the human consumes visualizations of the attributions and not the underlying numbers, we should take care that these visualizations do not distort the attribution in a way that undermines the justifications underlying the attribution methods. As Edward Tufte [25] notes:

*“Visual representation of evidence should be governed by principles of reasoning about quantitative evidence. For information displays, design reasoning must correspond to scientific reasoning. Clear and precise seeing becomes as one with clear and precise thinking.”*

<sup>1</sup>There are indeed applications of deep learning which take non-image inputs. For instance, a sentiment detection model like the one in [5] takes a paragraph of text as input. Consequently, the attributions of the score are to words in the text. In such cases, the attribution numbers can be communicated directly, one per word, without visualizations. Therefore the problems associated with vision applications do not arise, and we ignore such applications.



(a) Context surrounding the Attribution Visualizations.

**Figure 1: Two noteworthy points: (a) Visualization is a core component of the explanation module together with the algorithm that produces attributions. (b) The explanations are succeeded by the processes of interpretation and decision-making that are performed by the human. The overall effectiveness of the system depends on the prediction quality, the explanation quality, the human’s interpretation of the explanations, and the resulting decisions.**

The main purpose of this paper is to identify visualization principles that ensure that relevant properties of the attributions are truthfully represented by the visualizations. As we shall soon see, naive visualizations often only display a fraction of the underlying phenomena (Figure 2 (b)), and either occlude the underlying image or do not effectively display the correspondence between the image and the attributions (Figure 2 (e)).

### 1.3 Effectiveness of Visualizations

As we discuss visualization techniques, it is important to note that visualizations are used to help model developers debug data and models, and help experts to make decisions. In either scenario, the human has some competence in the model’s prediction task. For instance, doctors are capable of diagnosing conditions like diabetic retinopathy or mammography from images. In this case, the model and the explanation are either used to screen cases for review by a doctor (cf. [3]), or to assist the doctor in diagnosis (cf. [12]). The criterion for success here is the accuracy of the model and the human combined. To accomplish this, it is helpful to ensure that the humans comprehend the visualizations.

There are at least three reasons why the human may find it hard to understand the explanation: First, the model happens to reason differently from the human. Second, the attribution method distorts the model’s operation; this is a shortcoming of the attribution method and not the model. Third, the visualization distorts the attribution numbers; this is a shortcoming of the visualization approach.

While it is indeed possible to construct inherently interpretable models, or to improve the attribution method, the focus of the paper is solely on better visualizations. We will take the model behavior and the attribution method’s behavior as a given and optimize the visualizations.

Uncluttered visualizations tend to be easier to comprehend. Consider the difference between a scatter plot and a bar chart. The former displays a detailed relationship between two variables, but it can be relatively cluttered as a large number of points are displayed. In contrast, the latter is relatively uncluttered and reduces the cognitive load on the user. But it is possible that the binning (along the x-axis of the barchart) may cause artifacts or hide relevant information. Ideally, we’d like to reduce clutter without hiding information or causing artifacts.

If the visualization is cluttered the human may start ignoring the explanation altogether. This phenomenon is known as “disuse” of automation in the literature on automation bias [7]. However, if the visualization is over-optimized for human-intelligibility—for instance, by selectively highlighting features<sup>2</sup> that are considered important by the human—this could cause confirmation bias and suppress interesting instances where the model and the human reach different conclusions by reasoning differently. This phenomenon is known as “misuse” of automation. Both disuse and misuse ultimately harm the overall accuracy of the decision-making. Our goal, therefore, is to reduce clutter without causing confirmation bias.

A different aspect of successful comprehension of the explanations depend on how well the visualization establishes a correspondence between the two *layers* of information, namely the raw image and the depiction of the attributions. As Tufte [25] notes:

*“What matters—inevitably, unrelentingly—is the proper relationship among information layers.”*

Naively overlaying the attributions on top of the image may obscure interesting characteristics of the raw image. On the other hand, visualizing the two layers separately loses the correspondence between the explanation (the attributions) and the “explained” (the raw image). We would like to balance these potentially conflicting objectives.

## 2 GRAPHICAL INTEGRITY AND COVERAGE

Our first goal is to have the visualizations represent the underlying attribution numbers as faithfully as possible. The goal is to ensure that the visualizations satisfy the standard principle of **Graphical Integrity**, i.e., the visualization reflects the underlying data, or equivalently that the “representation of numbers should match the true proportions” (cf. [24]).

Intuitively, if a feature that has twice the attribution of another, then it should appear twice as bright. Though this is a fine aspiration, it will be rather hard to achieve precisely. This is because human perception of brightness is known to be non-linear [10] and it is also possible that there are spatial effects that affect human perception.

<sup>2</sup>Here, by feature we simply mean groups of logically related pixels, and not any formal notion of feature from either the computer vision or machine learning literature.

A corollary to Graphical Integrity is that features with positive or negative attributions are called out differently. This is easy to achieve by just using different colors, say green and red, to display positive and negative attributions. However, one can also naively translate feature “importance” in terms of high attribution *magnitude*, ignoring the sign of the attribution. This can be dangerous. As [19] discusses, the explanations can then appear to lack sensitivity to the network, i.e., the attributions from two very different networks can appear visually similar.

The most obvious way to achieve Graphical Integrity, assuming a calibrated display, is to linearly transform attributions to the range [0, 255]; the maximum attribution magnitude is assigned a value of 255 in a 8 bit RGB scheme.

Assume that this transformation is done separately for positive and negative attributions. Positive attributions are shown in green. Negative attributions in red.

As a concrete example, consider the image in Figure 2 (a), and an object recognition network built using the GoogleNet architecture [21] and trained over the ImageNet object recognition dataset [9]. The top predicted label for this image is “fireboat”; we compute attribution for this label<sup>3</sup> using the Integrated Gradients method [20] and a black image baseline. And we visualize only the pixels with positive attributions.

The resulting “naive” visualization (Figure 2 (b)) highlights some features that match our semantic understanding of the label: Primarily, attribution highlights a set of water jets on the left side of the image. Note, however, that the highlighted regions do not cover the entire area of this semantic feature: there are also water jets to the image right that are not highlighted, nor is the structure of the boat itself. Indeed, only a small fraction of the attributions is visible.

Such visualizations display only a small fraction of the total attribution. This is because the attributions span a large range of values, and are long-tailed (see Figure 3 (a)). We also show that this is not only true for Integrated Gradients, but also true for two other attributions techniques: Gradient x Image [14] and Guided Back-propagation [17] – showing that techniques from both categories (Section 1.1) of attribution exhibit this property. The large range implies that only the top few pixels<sup>4</sup> by attribution magnitude are highlighted (compare Figure 2 (b) against Figure 2 (c)). The long tail implies that the invisible pixels hold the bulk of the attribution. We found that in order to see 80% of the attribution magnitude, around 150000 pixels would need to be visible (out of 1 million pixels for a 1024x1024 image) across the three techniques (see Figure 3 (d)). In contrast, the naive visualization only shows (around) the top 500 pixels across the three techniques. It is possible that this long tail is possibly fixed by changes to either the deep network training process or the attribution, but this is outside the scope of this paper.

Let us revisit this phenomenon in our example. Suppose that we take only the pixels with positive attributions. Consider the pixel with the  $k$ th max attribution value such that the top  $k$  pixels

account for 75% of the attribution magnitude. Notice, that we would like this pixel to be visible so as to *cover* the underlying phenomena. Now suppose that we take the ratio of the max attribution value to  $k$ th max attribution value. This ratio is approximately 25. This would imply that if the max attribution value has a value close to 255 (in 8-bit RGB), the  $k$ th max pixel would have a value close to 10 and appear indistinguishable from black, and hence would be invisible.

This brings us to our second requirement, **Coverage**, which requires that a large fraction of important features are visible. This is a concept introduced by us, justified by the data analysis above; our other three requirements are standard either in the visualization or the computer vision literature.

To fix coverage, we must reduce the range of the attributions. There are several ways to do this, but arguably the simplest is to clip the attributions at the top end. This sacrifices Graphical Integrity at the top end of the range because we can no longer distinguish pixels or features by attribution value beyond the clipping threshold. But it improves coverage. We can now see a larger fraction of the attributions (see Figures 2 (d)). Back to our example. Suppose that we clip the attributions at the 99th percentile. The ratio of the maximum attribution value to that of the  $k$ th max pixel (such that the top  $k$  pixels account for 75% of the attribution magnitude) falls from 25 to 5. This would imply that if the max attribution value has a value close to 255, the  $k$ th max pixel would have a value close to 50, and this is distinguishable from black, and hence visible. Notice that we have achieved a high-degree of coverage (we can see 75% of the attributions) by sacrificing Graphical Integrity for a small fraction of the pixels (in this case, the top 1%).

So far we have visualized pixels with positive attributions. Certain pixels may receive a negative attribution. Strictly speaking, the Graphical Integrity and coverage requirements apply to visualizing such negative attributions too. However, empirically, we notice two different ways in which pixels with negative attribution arise. First, they coincide with pixels that have positive attribution as part of the same feature; this likely happens because the deep network performs edge detection and the positives and negatives occur on either side of the edge. Second, occasionally, we see entire features with negative attributions. In this case, the feature is a negative “vote” towards the predicted class. It is worth showing the second type of negative attribution, though possibly not the first type because they are redundant and simply increase cognitive load. For instance, in the case of the fireboat, pixels with negative attribution always co-occur with pixels with positive attributions; see Figure 2 (e).

In general, it is hard to know a priori that the negative attributions are simply redundant. We recommend first showing experts both positive and negative attributions initially, and suppressing the negative attributions if they are mostly redundant.

As another example, consider the scenario of explaining the diabetic retinopathy classification of a retinal image. This is arguably a more mission critical task than object recognition because it is health related. For the model, we use the deep learning model trained to detect diabetic retinopathy from [3], and again we use the Integrated Gradients method [20] for computing the attributions. Figure 4 shows a comparison between the naive and the clipped visualizations of the attribution.

<sup>3</sup>Specifically, we analyze the softmax value for this label. The softmax score is a function of the logit scores of all labels; explaining the softmax score implicitly explains the discriminatory process of the network, i.e., what made the model predict this label *and not any other label*

<sup>4</sup>Ideally we would like to report coverage numbers at the level of human-intelligible features. For analytical tractability, we will use pixels as proxies for features.

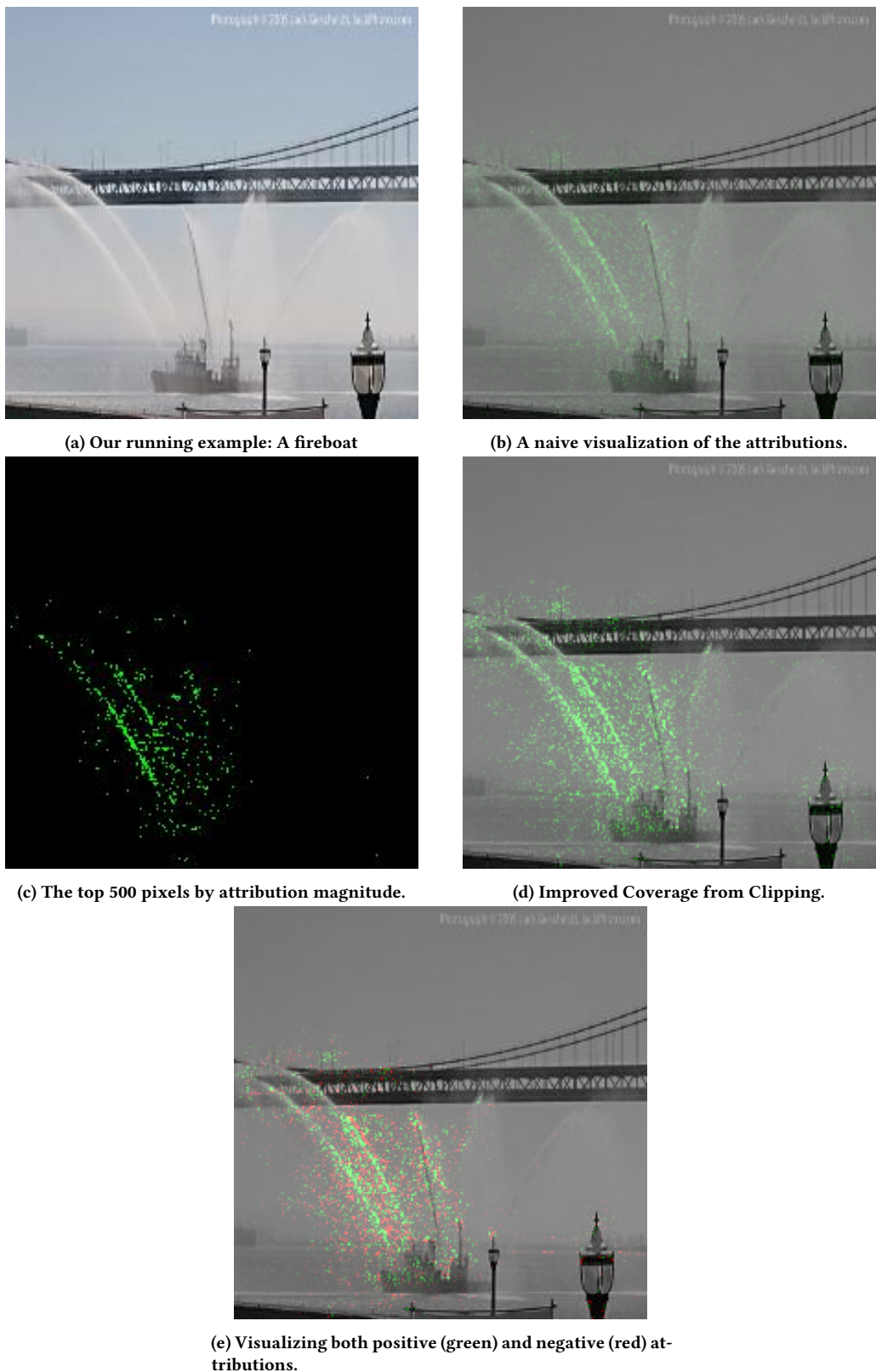
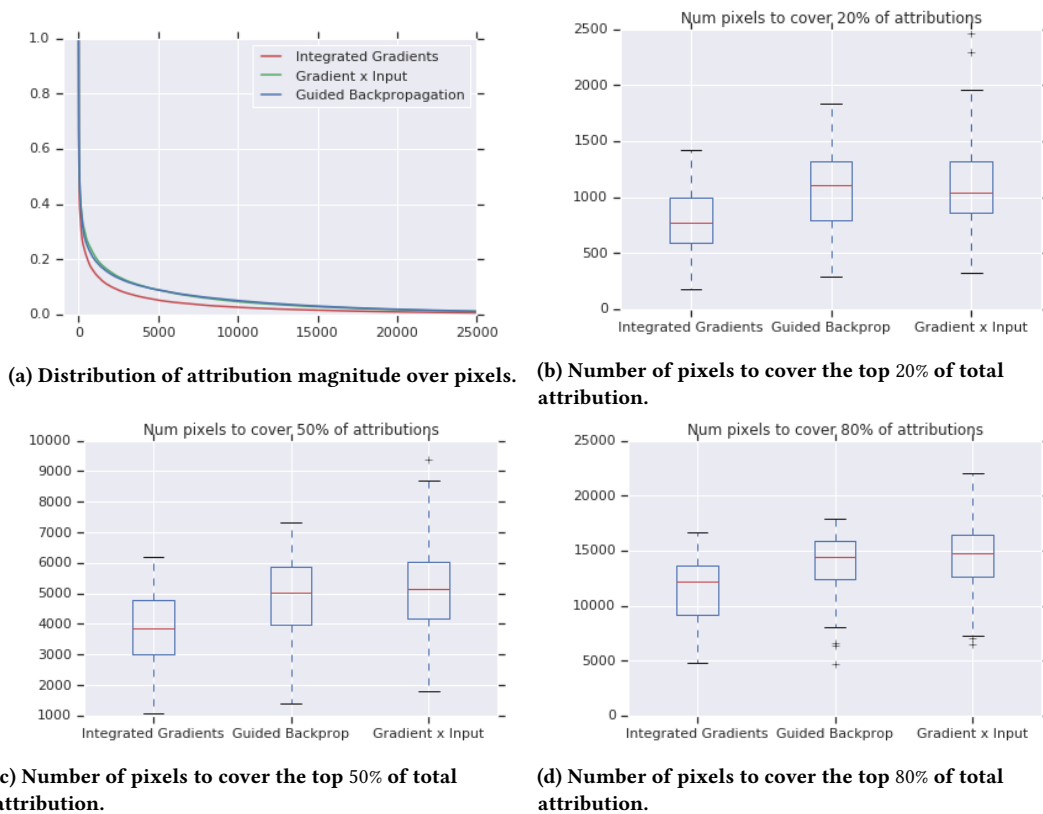


Figure 2: Naive visualizations have poor Coverage. (b) displays attributions that explain the prediction “fireboat” for the image in (a). (c) shows only the top 500 pixels by attribution. These pixels capture only 7.7% of the total attribution magnitude. Notice the similarity to (b). Together the two figures show that naive visualizations only display a small fraction of the attributions. (d) shows the increase in coverage from clipping the top of the range of the attributions. (e) extends (d) with visualizations. for negative attributions shown in red.



**Figure 3: Distribution of the attribution magnitudes. (a) shows the distribution of attribution magnitudes for the fireboat example (Figure 2). (b) (c) (d) show the top  $n$  pixels, sorted by attribution magnitude in descending order, to cover 20%, 50%, and 80% of total attribution, across 63 images.**

This example fundus image was determined by a panel of retina specialists to have severe non-proliferative diabetic retinopathy (NPDR), a form of the condition with increased risk of vision loss over time. This determination was made by the specialists based on image features indicating specific types of pathology: These include a range of microaneurysms and hemorrhages (dark regions) throughout the image, as well as potential intra-retinal microvascular anomalies (IRMA) that are diagnostic of more severe forms of the disease.

A deep-learning network accurately predicted Severe NPDR in this image; we examined the corresponding attribution map for this prediction. The naive visualization highlights some of this pathology, but misses much of it. Many microaneurysms and hemorrhages are omitted, particularly in the lower-right regions of the image (Figure 4 (a), arrows). The IRMA feature, important for accurate diagnosis, is somewhat highlighted, but of relatively low salience (Figure 4 (a), dotted circle). It is possible for clinicians to miss this signal.

By contrast, a clipped version of the image (which reduces the range of attributions by clipping attribution scores above a certain threshold; Figure 4 (b)) highlights these clinically-relevant features. We asked two ophthalmologists to evaluate this image, along with both visualizations, as part of our eval experiment (described in

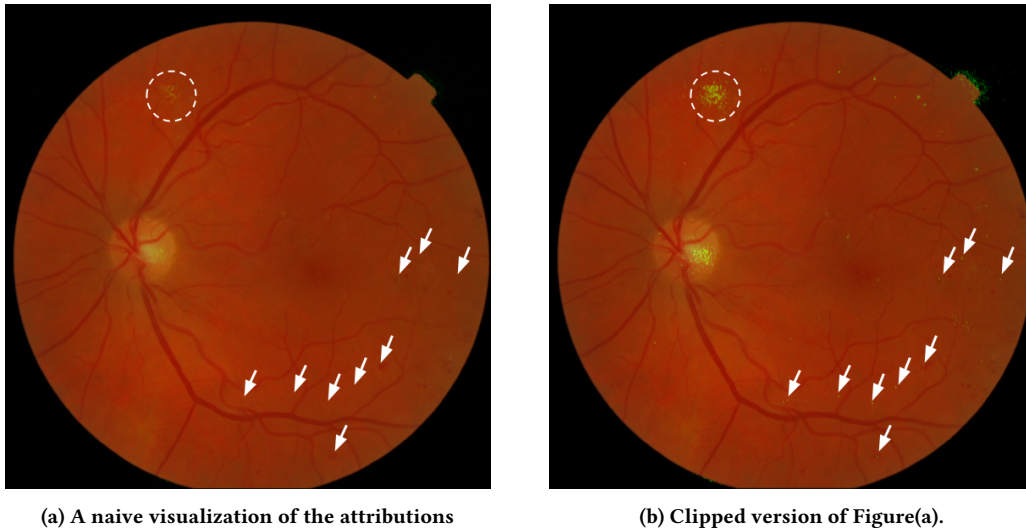
detail below). In this instance, both ophthalmologists indicated that they preferred the clipped version, citing the fact that the naive visualization either missed or inadequately highlighted the most important features for diagnosis.

### 3 MORPHOLOGICAL CLARITY AND LAYER SEPARATION

As we discussed in the Introduction (cf. Figure 1), attribution visualizations are ultimately for human consumption; a human interprets the visualization and takes some action.

Our next requirement is that the images satisfy **Morphological Clarity**, i.e., the features have clear form, the visualization is not “noisy”.

Notice that model may behave in a way that does not naturally result in Coherence; it could for instance, rely on a texture that is “noise” like. In this sense, optimizing for Morphological Clarity could be at the cost of faithfully representing the model’s behavior and the attributions. Nevertheless it is likely that visualizations that satisfy Coherence are more effective in an assistive context as we discussed in Section 1.3 and reduce cognitive load on the human. To best account for this trade-off, consider applying the operations below that improve Morphological Clarity as a separate knob.



**Figure 4: Explaining diabetic retinopathy classification.** Figure (a) displays attributions that explain the classification of proliferative diabetic retinopathy for a retinal image. Notice the lack of Coverage and the missing attributions to the hard exudates and microaneurysms outside the optic disk. Figure (b) shows a visualization with clipped attributions: Raw attribution scores below the 30th percentile value or above the 95th percentile value are clipped to the respective value. This version has greater sensitivity for highlighting clinically-relevant features.

To improve Morphological Clarity, We shall apply two standard morphological transformations. (a) The first fills in small holes in the attributions (called Closing) and (b) the second removes small, stray, noisy features from the visualization (called Opening). These are standard operations in image processing (cf. [13]). See Figure 5 for the result, where we applied the morphological transformations to both Integrated Gradients and Guided Backpropagation attributions. Notice that the morphological operations reduce visual clutter and therefore improve Morphological Clarity.

We now introduce our final requirement. Notice that our visualizations establish a correspondence between the attributions and the raw image, so that the attributions can highlight important features. It is important that the correspondence is established without occluding the raw image, because the human often wants to inspect the underlying image directly either to verify what the attributions show, or to form a fresh opinion. This is called **Layer Separation** (see Chapter 3 [23]), i.e., to ensure that both information layers, the attributions and the image, are separately visible.

Notice that we have been overlaying the attributions on top of a grayscale version of the original image. This ensures that the the colors of the attribution do not collide with those of the image and that we can associate the attributions with the underlying image. Unfortunately, we do not have a clear view of the raw image. If we had an interactive visualization, we could toggle Figures(e) and (a) on a click.

But a different alternative is to produce outlines of the important regions. This is also possible via standard morphological operations. The idea is to first threshold the attributions values into a binary mask, where all nonzero attribution values are mapped to 1, and the rest are zero. Then, we subdivide the mask into clusters by computing the connected components of the mask. Here, we can

rank the components by the sum of attribution weight inside each component, and keep the top N components. Finally, we can get just the borders around each kept component by subtracting the opening of the component by the component. See Figure 6. The result is that the underlying raw image is visible, but we can also tell which parts the attribution calls out as important.

#### 4 LABELING

An important aspect of the visualization is labeling. The labeling helps ensure that the decision-maker understands exactly what prediction is being explained. Deep networks that perform classification usually output scores/predictions for every class under consideration. The attributions usually explain one of these scores. It is therefore important to label visualizations with the class name and the prediction score. Additionally, it is worth clarifying whether a visualization distinguishes positive from negative attributions; and if so, indicate the sign of these attributions. See Figure 6 (d) as an example.

#### 5 EVALUATION

In order to evaluate the impact of the visualization principles discussed here on an actual decision-making task, we ran a pair of side-by-side evaluation experiments. For each experiment, 5 board-certified ophthalmologists assessed retinal fundus images for diabetic retinopathy (as in Figure 4), and then compared two visualization methods (such as Figure 4 (a) vs. Figure 4 (b)) in terms of how well they supported their evaluations<sup>5</sup>. In these comparisons,

<sup>5</sup>Since visualizations are inherently for human consumption, and since we have no source of ground truth, we decided on evaluating doctor preference as our ground truth. We selected a task for which doctors would be readily able to identify all clinically-relevant pathology relevant to diagnosis, reducing the chance of confirmation bias.



(a) Baboon, using Integrated Gradients attributions



(b) Baboon with morphological operations applied



(c) African Hunting Dog, using Integrated Gradients attributions



(d) African Hunting Dog with morphological operations applied

**Figure 5: Improving Morphological Clarity.**

the same underlying attribution map was used; the only thing that varied was the visualization parameters.

For the first comparison, we compared naive integrated gradients visualizations to a clipped version (where values above the 95th percentile are clipped to the 95th percentile value, and values below the 30th percentile are thresholded (set to zero)) (N=87 images). As we described in section 2, the clipped version sacrifices some Graphical Integrity for better Coverage. This allowed us to measure whether experts found this trade-off effective. For the second comparison, we compared the clipped version to an outline version, made using the methods we describe in Section 3 (N=51 images). This allowed us to measure the effect of the morphological operations to improve clarity. Due to doctor availability and time

Ideally, we should measure prediction accuracy of doctor aided with visualizations as in [12], but that is outside the scope of this paper.

constraints, we did not compare the naive version directly with the outline version.

For both experiments, each image and its resulting visualizations were independently assessed by 2 out of the 5 doctors participating in the experiment. For each image, doctors performed the following steps: (1) They viewed the original fundus image and were prompted “What is your assessment for diabetic retinopathy for this image?”. They could select from one of the 5 categories from the International Clinical Diabetic Retinopathy scale ([4]), plus a 6th option if the image was determined to be ungradeable by the doctor. (2) Doctors viewed an additional image, with the two visualizations being compared shown side-by-side. The assignment of visualization method to left/right side was randomized across trials. Doctors answered the following questions: “Which visualization better supports your evaluation of DR in this image?”; “(Optional) What contributed to your decision?”; “If you marked that you changed your diagnosis,

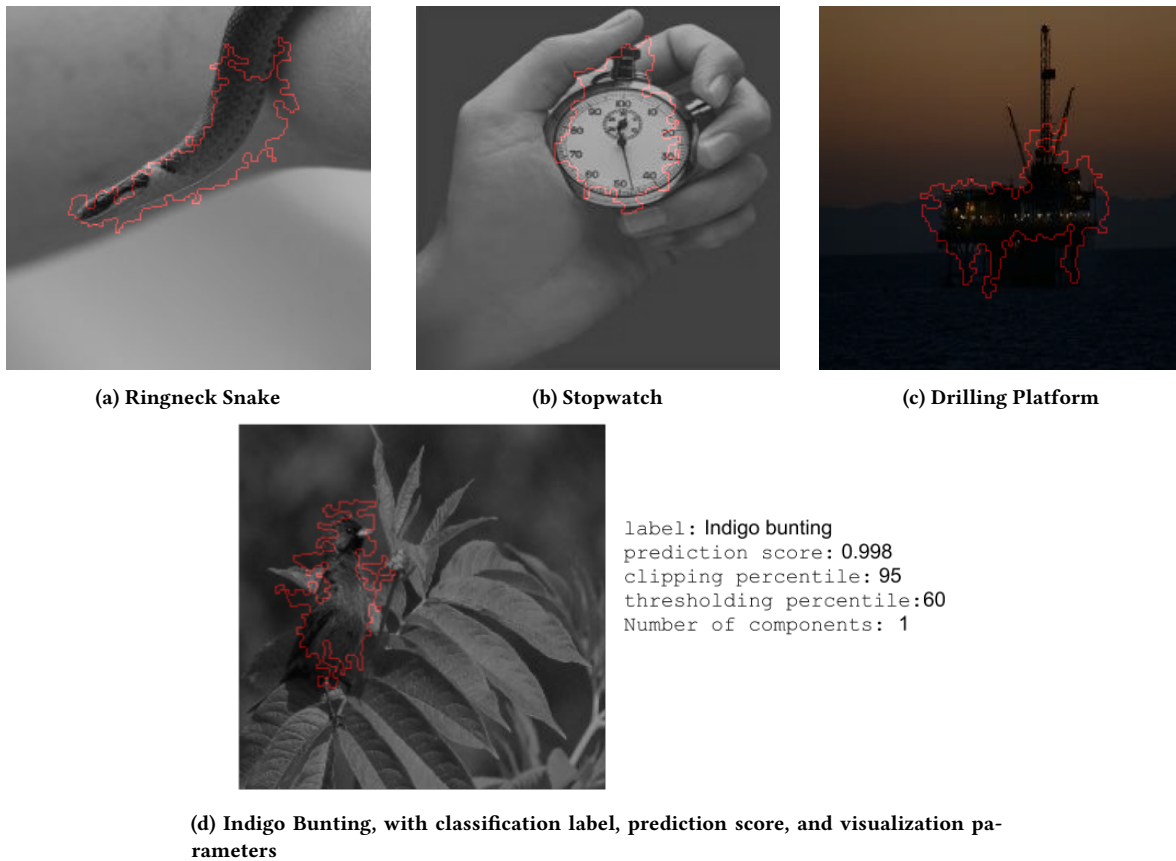


Figure 6: Outlines

what did you change it to?". For the second question, doctors could select 0 or more options from a range of reasons, including "The [left/right] visualization highlighted irrelevant features" and "The [left/right] visualization missed important features". We also included an option to indicate that the doctor changed their diagnosis after seeing the visualizations; and an "Other" option for qualitative open-text feedback on the value of the visualizations for the given image.

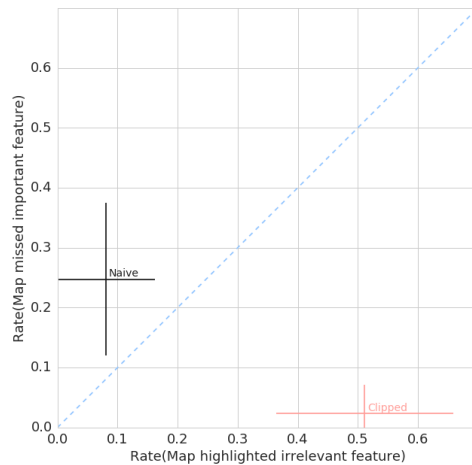
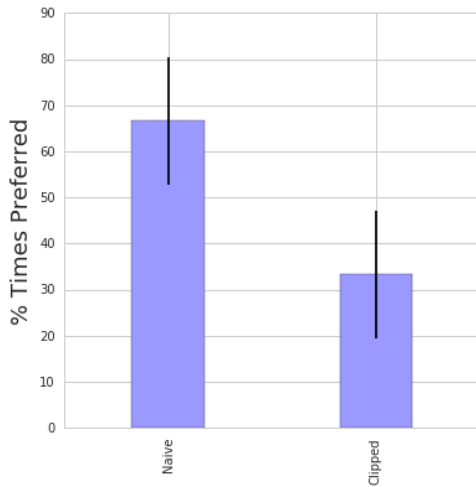
The results indicate that changes in Coverage and Morphological Clarity of a visualization can strongly impact its perceived value to consumers of the visualization (Figure 7). Overall, doctors tended to prefer a naive diabetic retinopathy visualization over our clipped version, at an approximately 2:1 rate (Figure 7 (a)). However, there were substantial differences between these versions in terms of the tradeoff between missing features and highlighting irrelevant features (Figure 7 (b)). Naive visualizations tended to miss important features more often (lacking sensitivity), while our clipped version tended to highlight irrelevant features (lacking specificity). Those instances when the naive visualization missed a feature tended to account for the approximately 1/3 of instances where a clipped version was preferred; however, given that a visualization covered the essential features, further increases in Coverage were not viewed as helpful, since the features revealed seemed less relevant. These

responses indicate that the clipping parameters used in this experiment were likely too aggressive at increasing Coverage. This indicates that, in this experiment, each visualization tended towards different errors, and that tuning Coverage (or allowing consumers of a visualization to tune Coverage, e.g. via a slider) can optimize this tradeoff.

We also found that increasing Morphological Clarity, via outlining, was preferred (Figure 7 (c)). Doctors tended to prefer outlines to the clipped version, again at an approximately 2:1 rate. This preference occurred despite the higher tendency of the outline version to miss features compared to the clipped version, (Figure 7 (d)). The relatively higher rate of sensitivity misses for outlines is likely due to the morphology step used here, which may remove some smaller features before computing the outlines. As with the Coverage experiment, the cases of sensitivity losses (when the visualization missed important features) account for many of the times when that outlines were not preferred (being cited in 31/36, or 86% of such trials); in cases where the outlines did not miss features, it was strongly preferred for increased clarity of the display.

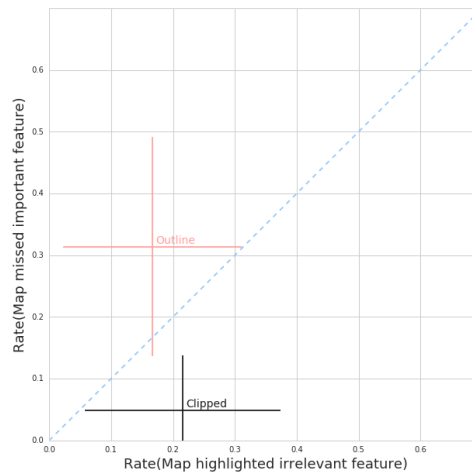
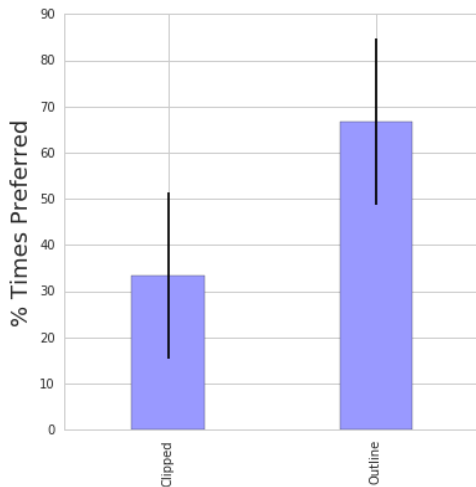
These experiments demonstrate the importance of clipping and Morphological Clarity to the potential utility of explanation visualizations in a clinical application. Adjusting parameters, such as clipping and outlining methodology, which affect a visualization's coverage and Morphological Clarity can play a strong role in





(a) Preference for two visualizations with different coverages with different coverage. Error bars reflect 95% binomial confidence intervals on the rate at which doctors evaluating the visualizations responded that a given nominal confidence intervals. Each doctor's assessment of visualization method highlighted irrelevant features (X axis) is treated as an independent observation. axis) or missed important features (Y axis)

(b) Tradeoffs between sensitivity and specificity of visualizations with different coverage. Conventions as in panel (a).



(c) Preference for two visualizations with different coherence levels on perceived effectiveness in explaining diabetic retinopathy predictions. Conventions as in panel (a).

(d) Tradeoffs between sensitivity and specificity of visualizations with different coverage. Conventions as in panel (a).

Figure 7: Explaining diabetic retinopathy classification. Results of side-by-side visualization evaluation experiments.

experts' preference for the visualization, even though the underlying attribution data are the same. They also affect whether the visualization highlights clinically relevant image features. These further illustrate that concepts such as sensitivity and specificity are applicable to the perception of visualizations. These tradeoffs may explain preference for a particular visualization type. They may also indicate the extent to which a human considers an explanation to diverge from their expected image features. (This latter signal may merit further exploration: When attribution highlights a feature that an expert considers irrelevant, this may indicate

model deficiencies, but it may also indicate features that a model has learned are diagnostic, but which diverges from the expected features that humans have learned to use.) Understanding the right balance of clipping and Morphological Clarity will be an important step in validating attributions to be useful for assisting people; and will likely depend strongly on the domain of the classification task. This may be an area where providing user control on aspects of a visualization may help it to be more effective.

## 6 CONCLUSION

We have evaluated our visualizations in the context of assisting doctors (see Section 5). In the introduction, we also mentioned a use case to help developers debug their models. In Figure 4 (b), notice that the model highlights the notch. The notch is not a pathology of diabetic retinopathy. It is perhaps the case that some camera makes had a predominance of certain types of DR cases, and the model picks up the notch as a predictive feature. This is obviously not desirable and highlights the ability of the visualizations to identify data and model issues.

There is a large literature on explaining deep-network predictions. These papers discuss principled approaches that identify the influence of base features of the input (or of neurons in the hidden layers) on the output prediction. However, they do not discuss the visualization approach that is used to *present* this analysis to the human decision-maker. The visualizations have a large influence on the effectiveness of the explanations. As we discuss in Section 5, modifying the visualization (by clipping it) change the types of model errors (sensitivity or specificity) detected by the human. As we discuss in Section 2, this difference is driven by the large range and heavy-tailedness of the attribution scores. Visualization is the language in which the explanations are presented and so it is important to treat it as a first-class citizen in the process of explanation, to be transparent about the visualization choices that are made, and to give the end user control over the visualization knobs.

Furthermore, we notice the central role of the human expert in the use of attribution maps within diabetic retinopathy diagnosis. The human implicitly *interprets* important logical/high-level features (e.g. a hemorrhage) from the pixel importances. In an assistive context, what matters is the prediction accuracy of the human and the model combined. We must rely on the human to be a domain expert, to be calibrated to perform the model’s (visual) prediction task, and to be calibrated to assess the visual explanations. We can tune the visualizations to aid the human in the process of interpretation. As discussed in Section 3, we can make the features in the visualizations more visually coherent and less noisy. Of course, we could go too far down this path and optimize for agreement between the human and the model; this would be dangerous as it would merely encourage confirmation bias.

Effective visualization will allow human experts to identify unintended or unexpected visual features, and relate their own understanding of the prediction task to that of the model’s performance. This will increase the combined accuracy of the model and the human.

Finally, we expect our visualization techniques to be applicable to almost all computer vision tasks. Real-world images and medical images capture a vast variety of tasks, and we demonstrate that our methods are applicable to both of these.

All our code is available at this link:  
<https://github.com/ankurtaly/Integrated-Gradients/tree/master/VisualizationLibrary>

## ACKNOWLEDGMENTS

We thank Naama Hammel, Rajeev Ramchandran, Michael Shumski, Jesse Smith, Ali Zaidi, Dale Webster, who provided helpful feedback and insights for this document.

## REFERENCES

- [1] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to Explain Individual Classification Decisions. *Journal of Machine Learning Research* (2010), 1803–1831.
- [2] Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. *CoRR* (2016).
- [3] V Gulshan, L Peng, M Coram, and et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316, 22 (2016), 2402–2410. <https://doi.org/10.1001/jama.2016.17216> arXiv:/data/journals/jama/935924/joi160132.pdf
- [4] S Haneda and H Yamashita. 2010. International clinical diabetic retinopathy disease severity scale. *Nihon rinsho. Japanese journal of clinical medicine* 68 (2010), 228.
- [5] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1746–1751. <http://aclweb.org/anthology/D/D14/D14-1181.pdf>
- [6] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4768–4777.
- [7] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39, 2 (1997), 230–253.
- [8] Doina Precup and Yee Whye Teh (Eds.). 2017. *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017*. Proceedings of Machine Learning Research, Vol. 70. PMLR. <http://jmlr.org/proceedings/papers/v70/>
- [9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* (2015), 211–252.
- [10] S S Stevens. 1957. On The Psychophysical Law. *Psychological review* 64 (06 1957), 153–81.
- [11] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Bach, and Klaus-Robert Müller. 2015. Evaluating the visualization of what a Deep Neural Network has learned. *CoRR* (2015).
- [12] Rory Sayres, Ankur Taly, Ehsan Rahimi, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, Shawn Xu, Scott Barb, Anthony Joseph, Michael Shumski, Jesse Smith, Arjun B. Sood, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2018. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology* (2018). <https://doi.org/10.1016/j.ophtha.2018.11.016>
- [13] Jean Serra. 1983. *Image Analysis and Mathematical Morphology*. Academic Press, Inc., Orlando, FL, USA.
- [14] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences, See [8], 3145–3153. <http://proceedings.mlr.press/v70/shrikumar17a.html>
- [15] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR* (2013).
- [16] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. *CoRR* abs/1706.03825 (2017). arXiv:1706.03825 <http://arxiv.org/abs/1706.03825>
- [17] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. 2014. Striving for Simplicity: The All Convolutional Net. *CoRR* (2014).
- [18] Yi Sun and Mukund Sundararajan. 2011. Axiomatic attribution for multilinear functions. In *12th ACM Conference on Electronic Commerce (EC)*. 177–178.
- [19] Mukund Sundararajan and Ankur Taly. 2018. A Note about: Local Explanation Methods for Deep Neural Networks lack Sensitivity to Parameter Values. <https://arxiv.org/abs/1806.04205>.
- [20] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks, See [8], 3319–3328. <http://proceedings.mlr.press/v70/sundararajan17a.html>
- [21] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. *CoRR* (2014).
- [22] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *CoRR* (2013).
- [23] Edward Tufte. 1990. *Envisioning Information*. Graphics Press, Cheshire, CT, USA.
- [24] Edward Tufte. 2001. *The visual display of quantitative information 2nd ed*. Graphics Press, Cheshire, Conn. <http://www.amazon.com/Visual-Display-Quantitative-Information/dp/0961392142>

- [25] Edward R. Tufte. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, Cheshire, CT.