Attribution in Scale and Space

Shawn Xu
Google AI Healthcare
jinhuaxu@google.com

Subhashini Venugopalan
Google Research
vsubhashini@google.com

Mukund Sundararajan
Google Inc.
mukunds@google.com

Abstract

We study the attribution problem [28] for deep networks applied to perception tasks. For vision tasks, attribution techniques attribute the prediction of a network to the pixels of the input image. We propose a new technique called Blur Integrated Gradients (BlurIG). *This technique has several advantages over other methods. First, it can tell at what scale a network recognizes an object. It produces scores in the scale/frequency dimension, that we find captures interesting phenomena. Second, it satisfies the scale-space axioms [14], which imply that it employs perturbations that are free of artifact. We therefore produce explanations that are cleaner and consistent with the operation of deep networks. Third, it eliminates the need for a 'baseline' parameter for Integrated Gradients [31] for perception tasks. This is desirable because the choice of baseline has a significant effect on the explanations. We compare the proposed technique against previous techniques and demonstrate application on three tasks: ImageNet object recognition, Diabetic Retinopathy prediction, and AudioSet audio event identification. Code and examples are on github*[1].

1. Introduction

There is considerable literature on feature-importance/attribution for deep networks [3, 28, 27, 4, 30, 21, 31, 20]. An attribution technique distributes the prediction score (*e.g.* sentiment or prevalence of disease) of a model for a specific input (*e.g.* paragraph of text or image) to its base features (*e.g.* words or pixels); the attribution to a base feature can be interpreted as its contribution to the prediction.

Suppose we have a model that predicts diabetic retinopathy (DR) from an image of the *fundus* of the eye. The attributions identify the (signed) importances of each pixel to the prediction. This tells us what part of the eye (*e.g.* retina/optic disc/macula), the network considers important to the prediction. We can use these attributions to debug the network—is the prediction based on a known pathology of diabetic retinopathy. The attributions could also assist the doctor. If the network's prediction differs from the doctor's the attributions could help explain why. This could improve diagnosis accuracy. [36] elaborates the application attributions to DR.

In this work, we study the attribution problem for perception tasks, i.e., the input is an image or a waveform and the features are pixels or time points. Perception tasks are different from others tasks (natural language, drug discovery, recommender systems): The base features (pixels or time points) are never influential by themselves; information is almost always contained in higher-level features like edges, textures or frequencies. This makes perception tasks worthy of separate study.

Our Contributions. We use the theory of scale-space [16] to make two contributions to the attribution literature:

Explanation in scale and space. Previous attribution techniques produced feature importance for pixels, *i.e.*, points in space. They do not produce localization in frequency. We propose a technique called Blur Integrated Gradients (BlurIG) to produce explanations in both scale/frequency and space. We can therefore tell that the detection of a steel-arch bridge is based on coarse, large-scale features, whereas the detection of a dog-breed depends on fine-grained features (see Figure 4b). (This is not simply a statement that dogs are smaller than bridges. In the dataset we study, the images frame dogs and bridges similarly.)

Perturbations free of artifacts. All attribution techniques involve perturbations of either the inputs (e.g. [31]) or the network state (e.g. [28]). The premise is that *removing* an important feature causes a large change in the prediction score. The literature does not however discuss if these perturbations could accidentally *add* features. Then, the change in the score could be because a different object is detected, and not because information is destroyed. This could result in the explanation artifacts that identify influential features that are not actually present in the input. (Compare the IG and Blur IG explanations for 'starfish' in Figure 3.) We discuss how the scale space-theory addresses accidental feature creation.

[1] https://github.com/PAIR-code/saliency

2. Related Work

There is considerable literature on feature-importance/attribution for deep networks, i.e. techniques to identify the influence/importance of each feature to the prediction of a network for a specific input. Some techniques perform this attribution by propagating the prediction from the output back towards the input (*e.g.* [28, 20, 27, 29]). Other techniques build upon the gradient of the prediction with respect to the input or a coarse version of the input (*e.g.* [26, 31, 11]). Our work implicitly addresses a critique of Integrated Gradients identified in [11]—that the explanations depend on the auxiliary baseline parameter, and varying the baseline changes the explanation. [11] addresses this criticism by averaging the attributions over two baselines (a black and a white image). In contrast the technique we propose subsumes the baseline; there is no need for this parameter.

There are also other techniques like [24] that are not specific to deep networks. Another class of techniques uses gradient ascent to modify the input to emphasize features critical to a neuron activation, either for an output neuron, or an internal one [22]. [5] formulates an optimization problem to highlight regions of an image that are important to a certain prediction. They incorporate a smoothness criterion based on Gaussian blur within the optimization objective to produce smooth, contiguous masks. Though the explanations have higher verisimilitude for real-world images, it is possible that they do not pick up fine features or textures, aspects that are critical to medical diagnosis applications (*e.g.* Section 5.2).

None of the works above address attribution along the scale/frequency dimension, or ensure that the perturbations don't manifest new information. We borrow concepts from the literature on scale-space theory [12, 14, 37]; this is a theory of multi-scale signal representation that represents the an image or signal as a one-parameter family of smoothed images or signals. We use the prediction gradients of this family of images to produce explanations. The smoothing process is axiomatically guaranteed *not* to produce artifacts in the image(see Section 4.1) using the so called scale-space axioms [14]. Therefore the resulting explanations will also be free of artifacts. There is a large literature that uses scale-space theory and Gaussian blurs to detect blobs, edges or other higher-level features *e.g.* [19, 15]; this literature is unrelated to deep networks or the interpretability/explanation/feature importance problem.

Finally, we mention a brief connection to cooperative game theory. Integrated Gradients [31] that we build on, is itself based on a method called Aumann-Shapley [2]. The method is constructed axiomatically using axioms that characterize desirable properties of attribution techniques *e.g.* [7]. These axioms differ from the scale-space axioms mentioned earlier. The scale-space axioms axiomatize the type of perturbation.

3. Attribution in Scale and Space

3.1. Blur Integrated Gradients (BlurIG)

BlurIG extends the Integrated Gradients [31] technique. Formally, suppose we have a function $F : \mathsf{R}^{m\times n} \rightarrow [0,1]$ that represents a deep network. Specifically, let $z(x,y) \in \mathsf{R}^{m\times n}$ be the 2D input at hand, and $z'(x,y) \in \mathsf{R}^{m\times n}$ be the 2D baseline input, meant to represent an informationless input. For vision networks, this could be the black image, or an image consisting of noise. We consider the straight-line path (in $\mathsf{R}^{m\times n}$) from the baseline z' to the input z, and compute the gradients at all points along the path. The path can be parameterized as

$$\gamma(x,y,\alpha) = z'(x,y) + \alpha \cdot (z(x,y) - z'(x,y))$$

Integrated gradients (IG) are obtained by accumulating these gradients. The integrated gradient for an input z and baseline z' is:

$$\mathsf{IG}(x,y) ::= (z(x,y) - z'(x,y)) \cdot \int_{\alpha=0}^{1} \frac{\partial F(\gamma(x,y,\alpha))}{\partial \gamma(x,y,\alpha)}\, d\alpha \quad (1)$$

Let us call this the *intensity scaling* IG; if you use a black image as the baseline z', the path scales the intensity of the image. Let us instead consider the path defined by successively blurring the input by the Gaussian blur filter. Formally, let

$$L(x,y,\alpha) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \frac{1}{\pi\alpha} e^{-\frac{x^2+y^2}{\alpha}} z(x-m, y-n)$$

be the discrete convolution of the input signal with the 2D Gaussian kernel with variance α, also known as the scale parameter. Blur integrated gradients is obtained by accumulating the gradients along the path defined by varying the α parameter:

$$\mathsf{BlurIG}(x,y) ::= \int_{\alpha=\infty}^{0} \frac{\partial F(L(x,y,\alpha))}{\partial L(x,y,\alpha)} \frac{\partial L(x,y,\alpha)}{\partial \alpha}\, d\alpha \quad (2)$$

Implementation-wise, the integral can be efficiently approximated using a Riemann sum:

$$\mathsf{BlurIG}(x,y) \approx \sum_{i=1}^{s} \frac{\partial F(L(x,y,\alpha_i))}{\partial L(x,y,\alpha_i)} \frac{\partial L(x,y,\alpha_i)}{\partial \alpha_i} \frac{\alpha_{\text{max}}}{s}$$

where $\alpha_i = i \cdot \frac{\alpha_{\text{max}}}{s}$ and s is the number of steps in the Riemann approximation. The gradients are obtained using numerical differentiation. The maximum scale α_{max} should be chosen to be large enough so that the resulting maximally blurred image is information-less. Small features are destroyed at smaller scales. But the specific value of α at which a feature is destroyed depends on the variance of the intensity of the pixels that form the feature (see Figure 5.10 from [14] for a discussion of this). The smaller the variation, the smaller the α at which it is destroyed.

Note that we have defined blur integrated gradients for 2D signals, but it exactly equivalent (after reparameterization) for the 1D case.

3.2. Interpretation of Blur IG

It is well known that $L(x, y, \alpha)$ in equation 2 is the solution of the 2D diffusion equation:

$$\frac{\partial L}{\partial \alpha} = \frac{1}{4}\nabla^2 L$$

with the initial condition $L(x, y, 0) = z(x, y)$. By commutativity of differentiation with convolution,

$$\nabla^2 L = \nabla^2 (G * z) = (\nabla^2 G) * z$$

where $G(x, y, \alpha)$ is the Gaussian kernel and $\nabla^2 G(x, y, \alpha)$ is the Laplacian of Gaussian (LoG) kernel. Thus, equation 3 can be alternatively expressed as:

$$\mathsf{BlurIG} ::= \frac{1}{4}\int_{\alpha=\infty}^{0} \frac{\partial F(L)}{\partial L} \cdot (\nabla^2 G) * z \, d\alpha \qquad (3)$$

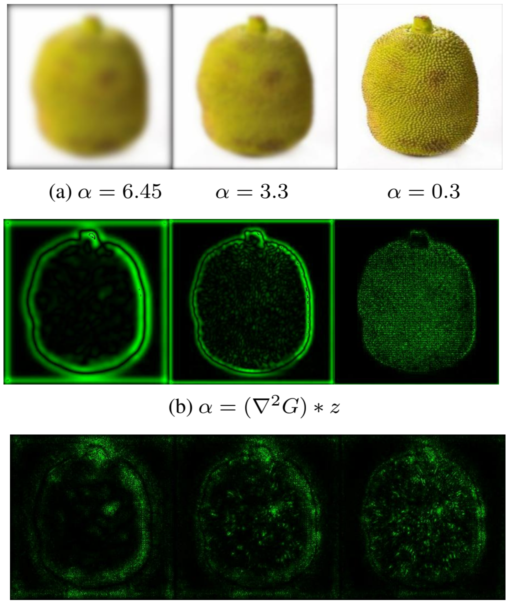
The LoG kernel is a band-pass filter used to detect edges at the specified scale. Thus, the first term of the integrand, which consists of model gradients with respect to the blurred image at the appropriate scale provides localization in space. This when filtered by the LoG-filtered image provides localization in frequency.

Figure 1 shows the different components of the integrand at different scales along the blur path while explaining the 'jackfruit' prediction for the Inception-Resnet-v2 model trained on ImageNet. At low resolution, the model picks up high-level details of the jackfruit, such as the stem and overall shape, while at high resolutions, the model picks up the spikiness of the texture.

4. Applying Scale-Space Axioms to Attribution

Attributions and Explanations are based on perturbations. Every method—for instance, Gradcam, IG, or Blur IG —prescribes a specific set of perturbations to the input (*e.g.* gradient computation). If the perturbations destroy 'information', then the resultant change in prediction can be interpreted as feature importance; this is the desired interpretation. However, if the perturbation creates information, then the resultant change in score is not due to a feature present in the input, and the result will be a misleading, uninterpretable explanation.

In this section, we use the scale-space theory [37, 12] to identify sequences of input perturbations that provably do not manifest information.



(a) $\alpha = 6.45$ $\alpha = 3.3$ $\alpha = 0.3$

(b) $\alpha = (\nabla^2 G) * z$

(c) partial Blur IG

Figure 1: Aspects of Blur IG at three levels of Blur (α). First row is the input. Second row is the Laplacian. Third row is the Blur IG integration (summation) up to that α.

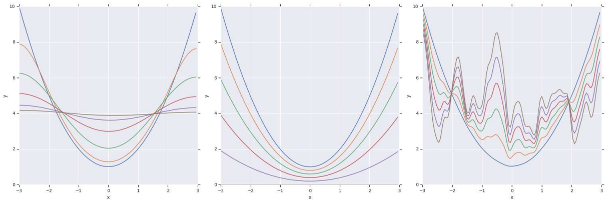
4.1. The Causality Axiom

Consider a sequence of grayscale images $L : R^3 \to R$; here $L(x, y, t)$ denotes the intensity of pixel x, y at scale t. The input image is $L(x, y, 0)$. As the scale parameter t increases, so does the level of perturbation. Scale-space theory formalizes 'Non-Creation of New Structure'[14] using the following 'Causality' axiom[12]:

Causality: Local extrema should not be enhanced as you increase the scale parameter. Namely, if $L(x, y, t)$ is at a local maxima (resp. minima) in x, y, then L should not increase (resp. decrease) with t. Equivalently, no new level surfaces are created as the scale parameter increases. Visual features like edges or textures correspond to local extrema in the representation $L(x, y, \cdot)$. Therefore, the causality axiom states that features should only be destroyed as t increases. In fact, the axiom states something stronger: that existing extrema are not enhanced by the path.

Figure 2 shows the scale space for blur scaling and intensity scaling (both black baseline and random baseline) for the univariate function $x^2 + 1$. Blur satisfies causality—no new extrema are created and the only minima is diminished. Intensity scaling (black baseline) breaks the stronger form of causality—the only minima is enhanced. Intensity scaling (random baseline) breaks the weaker form of causality as well – not only can the minima be diminished, but new extrema are introduced.

Remark 1 (Baseline choice in IG) *If we use a black image as a baseline (as suggested by [31]) for IG, we get a series of images that differ in intensity. As Fig. 2b shows, min-*



(a) Gaussian blur (b) Intensity scaling (black baseline) (c) Intensity scaling (random baseline)

Figure 2: Scale space for $x^2 + 1$ along the Gaussian blur, and intensity scaling (black and random baseline) paths.

ima can be enhanced along this path. However it is easy to see that no new extrema are created by intensity scaling. In this sense, it satisfies a weaker notion of causality axiom. However, as discussed by [11], using the black image as baseline suppresses intuitively important features that have low intensity. One remedy is to use a noise image as a baseline, as is done in [1]. However this can create new extrema (Fig. 2c), that result in spurious features (the 'starfish' example from Fig. 3 is an instance of this.) This is somewhat mitigated by averaging the attributions over several runs of the method. But this is a heuristic fix, and involves a computational hit. Blur IG appears to be a more systematic fix.

4.2. Justifying Blur Integrated Gradients

4.2.1 Path Methods

IG and Blur IG aggregate the gradients along a series of images that culminate at the input. Clearly, there are many other possible paths, and each such path will yield a different attribution method. We first justify why path methods are desirable.

Formally, let $\gamma = (\gamma_1, \dots, \gamma_n) : [0, 1] \rightarrow \mathsf{R}^n$ be a smooth function specifying a path in R^n from the baseline z' to the input z, i.e., $\gamma(0) = z'$ and $\gamma(1) = z$. Given a path function γ, *path integrated gradients* are obtained by integrating the gradients along the path $\gamma(\alpha)$ for $\alpha \in [0, 1]$. Formally, path integrated gradients for an input z is defined as follows.

$$\mathsf{PathIntGrads}^{\gamma}(z, z') ::= \int_{\alpha=0}^{1} \frac{\partial F(\gamma(\alpha))}{\partial \gamma(\alpha)} \frac{\partial \gamma(\alpha)}{\partial \alpha} \, d\alpha \qquad (4)$$

Attribution methods based on path integrated gradients are collectively known as *path methods*. Notice that integrated gradients is a path method for the straightline path specified by $\gamma(\alpha) = z' + \alpha \cdot (z - z')$ for $\alpha \in [0, 1]$. More interestingly, path methods are the only methods that satisfy certain desirable axioms. (For formal definitions of the axioms and proof of Proposition 1, see Friedman [7].)

Axiom: Dummy. If the function implemented by the deep network does not depend (mathematically) on some variable, then the attribution to that variable is always zero.

Axiom: Linearity. Suppose that we linearly composed two deep networks modeled by the functions f_1 and f_2 to form a third network that models the function $a \cdot f_1 + b \cdot f_2$, *i.e.*, a linear combination of the two networks. Then we'd like the attributions for $a \cdot f_1 + b \cdot f_2$ to be the weighted sum of the attributions for f_1 and f_2 with weights a and b respectively. Intuitively, we would like the attributions to preserve any linearity within the network.

Completeness if for every explicand z, and baseline z', the attributions add up to the difference $f(z) - f(z')$, the difference in prediction score for the input and the baseline.

Affine Scale Invariance (ASI) if the attributions are invariant under a simultaneous affine transformation of the function and the features. That is, for any c, d, if $f_1(z_1, \dots, z_n) = f_2(z_1, ..., (z_j - d)/c, ..., z_n)$, then for all j we have $\text{attr}_j(z, z', f_1) = \text{attr}_j((z_1, \dots, c * z_j + d, \dots z_n), (z'_1, \dots, c * z'_j + d, \dots z'_n), f_2)$, where z_j is the value of the j-th pixel. ASI conveys the idea that the zero point and the units of a feature should not determine its attribution; this is defensible for machine learning where the signal is usually carried by the covariance structure of the features and the response variable.

Proposition 1 *(Theorem 1 [7]) Path methods are the only attribution methods that always satisfy Dummy, Linearity, Affine Scale Invariance and Completeness.*

We now justify why the path used by Blur IG is superior to other paths for perception tasks using scale-space theory. Let $z(x, y)$ be the signal and $k(x, y, t)$ for all $t > 0$ be the family of infinitely differentiable, rapidly decreasing kernels such that the discrete convolution:

$$L(x, y, t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} k(m, n, t) z(x - m, y - n)$$

represents the path for $t > 0$, with the initial condition $L(x, y, 0) = z(x, y)$. Kernels are linear, shift-invariant. A kernel is symmetric if $k(x, y, t) = k(-x, y, t)$ and $k(x, y, t) = k(y, x, t)$. Symmetry ensures that the transformation is identical in every direction in the x, y plane. A kernel satisfies the semigroup property if $k(\cdot, \cdot, t_1) * k(\cdot, \cdot, t_2) = k(\cdot, \cdot, t_1 + t_2)$. The semigroup property ensures that all scales are treated identically. We then have the following proposition

Proposition 2 *(Theorem 3.2 [14]) The only kernel method that is symmetric, satisfies the semi-group property, satisfies continuity in the scale parameter, and causality, is the Gaussian kernel.*

Remark 2 *We briefly contrast the axiomatization of IG from [31] to this axiomatization of Blur IG. The axiomatization for IG built on top of Proposition 1. It used an additional axiom of Symmetry, that variables symmetric in* *the function, with equal value in the input and the baseline get identical attribution. This is a condition on the* function *implemented by the deep network. Notice that the condition says nothing about the location (in x, y) of the two features. The condition only appears to make sense where there is no 'geometry' to the input, for instance when the task is a natural language one. In contrast, Blur IG uses Proposition 2 (in addition to Proposition 1), which uses axioms about the transformations that generate the path of images. These conditions on the transformation crucially rely on the geometry of the input. for instance the condition that makes a kernel symmetric, or what makes a point a extrema in the definition of causality.*

5. Applications

5.1. Object Classification

We apply Blur IG to the Inception-Resnet-V2 architecture [35] trained on the ImageNet classification challenge [25]. Figure 3 shows the comparison of three different explanation techniques – GradCAM [26], IG [31] (random baseline), and Blur IG – applied to various example images. Notice that GradCAM sometimes produces an attribution mask too coarse in resolution to capture the morphology of the object (*e.g*. windsor, starfish in Fig. 3), and random baseline IG sometimes results in spurious artifacts outside of relevant regions (*e.g*. starfish, fur coat in Fig. 3); Remark 1 discusses this issue. In the last row of Figure 3 for a container correctly predicted as 'eggnog', only Blur IG attributes the packaging label correctly, while GradCAM and IG attribute the top portion of the container. It is of course possible that the top portion is sufficient to discriminate 'eggnog'. However, we found that the model classifies images of containers of similar shape and color, but without the packaging label, as 'water jug'. Therefore it appears that the critical feature, the packaging label, is missed by the techniques, and captured by Blur IG.

One application of Blur IG is it's ability to answer the question: **At what scales does the network recognize the image features relevant for the predicted class?** Figure 4 compares the relative scales at which 'steel arch bridge' and 'maltese dog' are classified. We can see from Figure 4b that the bulk of the Blur IG attributions occur from $\sigma = 6$ to $\sigma = 3$ for 'steel arch bridge', while Figure 4a shows that it is from $\sigma = 3$ to $\sigma = 0$ for 'maltese dog'. Thus the dog prediction happens at far lower scales; this is probably because dog species are classified based on fine-grained features. Note that in the ImageNet data set, the objects are usually in focus, i.e., the dogs appear as large as the bridges. See figure 5. So we are not simply saying that dogs are smaller than bridges.

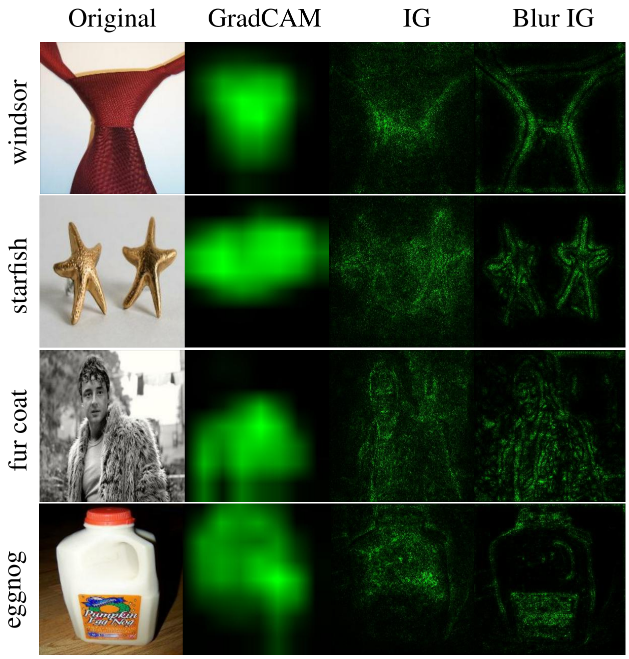
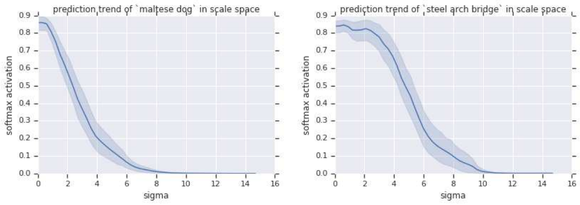
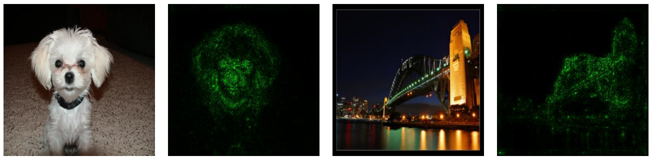


Figure 3: Comparison of GradCAM, random baseline IG, and blur IG on a sample of images from ImageNet .



(a) 'maltese dog' trend (b) 'steel arch bridge' trend

Figure 4: Comparison of the prediction trends of 'maltese dog' vs 'steel arch bridge' along the blur integration path. The bulk of the prediction weight for 'maltese dog' accumulates between $\sigma = 0$ and $\sigma = 3$, whereas the bulk of the prediction weight for 'steel arch bridge' accumulates between $\sigma = 3$ and $\sigma = 6$. This matches our hypothesis that the model can recognize the bridge at course scale/low frequency, whereas it requires finer scale / high frequency details to recognize the dog breed (from other breeds).



(a) maltese dog (b) Blur IG (c)steel arch bridge (d) Blur IG

Figure 5: Example images of 'maltese dog' and 'steel arch bridge' along with their Blur IG attribution masks.

Remark 3 (Visualizations) *For perception tasks, the attributions are almost always communicated as pictures; the scores are not directly communicated. To inspect the frequency/scale dimension, visualizing a series of saliency maps (e.g. Figure 1) is preferable to standard saliency maps (e.g. Figure 3). The quality of the explanations do depend on the quality of visualizations [32].*

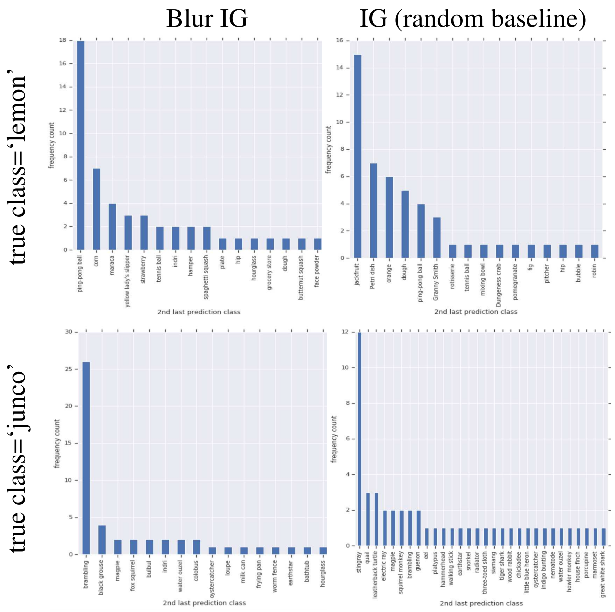


Figure 6: Histograms (N=50) of 2nd last predicted class along integration path of Blur IG (left) vs IG (right) for the classes 'lemon' (top row) and 'junco' (bottom row).

Blur IG path produces more natural perturbations. To test this, we sampled 100 classes, with 50 true positive images per class, and for each image found the second last label on the integration path that is not equal to the true class. If the perturbation were natural, we would expect the second last label to be something that a human would find a plausible alternate classification. See Figure 6 for examples of second last class for images from two classes, 'lemon' and 'junco' (a bird). The most frequent 2nd last label for IG, 'jackfruit' is semantically closer to 'lemon' than the most frequent 2nd last class for Blur IG, 'ping pong ball'. However, a ping-ping ball is morphologically similar to a lemon. For 'junco', IG's most frequent 2nd last label, 'stingray', seems unrelated semantically and morphologically.

Quantitative evaluation using ImageNet hierarchy. To make the analysis objective, we use WordNet. WordNet imposes an ontology (tree) on top of the ImageNet labels. The ontology (tree) mirrors human intuition. The parent of 'maltese' is 'dog' and its parent is 'animal' and so on. Classes that have a closer tree distance are intuitively more related. We compare the WordNet tree distance between the second last label and the true label for Blur IG, IG (random baseline), and IG (black baseline), where tree distance is defined as the sum of the distances to the lowest common ancestor of the second last label and the true label. Our analysis shows that Blur IG has the lowest average distance at 8.59, while IG with random baseline has a larger distance of 9.17, confirming that Blur IG employs more natural perturbations. To test for statistical significance, we used a paired-sample t-test with $H_0 : \mu_D <= 0.25, H_1 : \mu_D > 0.25$, where μ_D represents the mean difference between blur and IG random baseline label distances. With confidence level 95%, we reject the null hypothesis with $t = 1.674 > t_{0.05,99} (= 1.6604)$. Figure 6 (bottom row) compares 2nd last label histograms for the 'lemon' class, for which the IG path has a lower average WordNet tree distance. We hypothesize that the improvement in quality of the explanation masks produced by Blur IG is a result of it having a more 'natural' integration path from the counterfactual input to the real input, compared with the IG integration path (see Remark 1, and Section 6 for a further discussion of this).

Quantitative evaluation using human segmentation. Next, we perform quantitative evaluations proposed in [11]. For ImageNet we use the Inception-v1 model as in [11]. Given an annotation region, the evaluation computes AUC, F1, and MAE (mean absolute error) of the generated saliency mask by considering pixels within the annotation to be positive and outside to be negative. In our evaluations, the annotations correspond to bounding-boxes for ImageNet, and segmentation masks for Diabetic Retinopathy (DR) (Sec. 5.2), we report the average scores in Table 1.

Attribution method	ImageNet			Diabetic Retinopathy		
	AUC ↑	F1 ↑	MAE ↓	AUC ↑	F1 ↑	MAE ↓
XRAI	0.836	0.786	0.149	0.805	0.285	0.068
GradCAM	0.742	0.715	0.194	0.817	0.249	0.058
IG (random-4)	0.709	0.674	0.223	0.827	0.344	0.060
IG (black)	0.710	0.674	0.219	0.828	0.307	0.062
IG (black+white)	0.729	0.681	0.216	0.818	0.296	0.062
Blur IG (ours)	0.738	0.693	0.209	0.831	0.293	0.061

Table 1: Average F1, AUC, and MAE scores for different explanation methods on images from ImageNet validation set (N=9684), and Diabetic Retinopathy dataset (N=141). (↑ indicates higher is better, ↓ indicates lower is better)

GradCAM and XRAI outperform BlurIG on Imagenet, but BlurIG does better on the DR task. There's a difference between the methods on the two tasks since XRAI (and GradCAM) have lower resolution explanations. XRAI employs image segmentation (that is oblivious to the model), and then groups the IG (Black+White) attributions by these segments. This trades faithfulness to the model's behavior for better visual coherence on natural images (ImageNet), but makes the results worse for tasks like DR that do not involve "natural" features. Finally, we note that the segmentation idea is orthogonal to the attribution technique, i.e., XRAI could use BlurIG instead of IG (Black+White).

Note on biases. Different methods have different biases. We note that IG has a bias towards color depending on the color contrast (or the lack thereof) between the class of interest and the baseline whereas BlurIG has a bias towards shape. This is illustrated in Fig. 7 below.

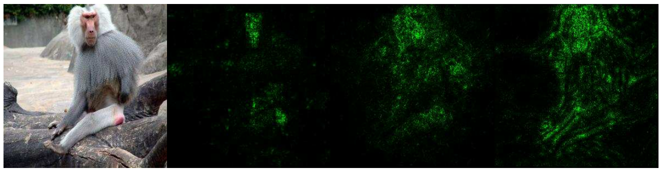

(a) Original (b) IG (grayscale) (c) IG (black) (d) Blur IG

Figure 7: [Bias] Examples to show that IG has bias towards color and Blur IG has a bias towards shape. The 4 images (left to right) correspond to the original, followed by saliency masks for IG with a grayscale version of the input image as baseline, IG with a black baseline, and Blur IG. Notice that IG (grayscale) focuses on the baboon's pink face and bottom, and IG (black) emphasizes light features. In contrast, Blur IG emphasizes its overall shape, paying close attention to the shape of the head and limbs.

5.2. Diabetic Retinopathy

We also apply Blur IG to a Diabetic Retinopathy prediction task. Specifically, we study the model from [36] that uses the Inception-V4 architecture [33]. Figure 8 compares GradCAM, random baseline IG, and Blur IG on retina images diagnosed with diabetic retinopathy (DR). The first image contains hard exudates (small white or yellowish deposits), retinal hemorrhages (dark spots), and microaneurysms (small, round, dark spots), with a diagnosis of moderate DR determined by retina specialists. Notice that the GradCAM explanation is too coarse-grained to be appropriate for explanation of DR classification. The random baseline IG explanation correctly attributes the hard exudates at the left and top of the image, but misattributes the bright spot on the optic disk, which is a result of a camera artifact (brightness oversaturation). In contrast, the Blur IG explanation correctly attributes the DR lesions and ignores the bright spot. The second image contains an example retina image diagnosed with proliferative DR. Critical to this diagnosis is the presence of neovascularization of the optic disc (NVD). GradCAM fails to attribute the NVD lesion (its attribution in that region is purely negative, which indicates that the lesion is not relevant to the classification of DR). Both random baseline IG and Blur IG attribute the NVD, but random baseline IG also seems to highlight numerous spurious spots that does not seem to have clinically relevant features. In addition, Blur IG attributes blood vessels that possibly contain venous beading, another lesion associated with proliferative DR.

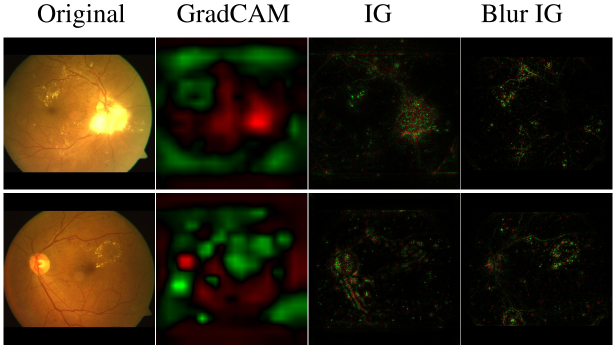


Figure 8: Comparison of GradCAM, random baseline IG, and Blur IG on retina images diagnosed with diabetic retinopathy. Green and red colors indicate positive and negative attributions respectively.

5.3. Audio Classification

We apply Blur IG to study a CNN model [10] trained on AudioSet [8] audio event recognition, i.e., the task of predicting types of audio events from an ontology of 635 classes that range across human sounds, animal sounds, musical sounds, sounds made by objects, etc. The CNN is a modified ResNet-50 [9]. The model takes the spectrogram of the waveform. (A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time.) The audio is processed as non-overlapping 960ms frames. These are decomposed with a short-time Fourier transform applying 25 ms windows every 10 ms. The resulting spectrogram is integrated into 64 mel-spaced frequency bins, and the magnitude of each bin is log transformed. This gives log-mel spectrogram patches of 96X64 bins that form the input. The output of the model is a multilabel classification. For instance, an audio sample of a violin is expected to be classified as a violin but also as a bow-stringed instrument, and as music. We apply this model on audio samples publicly available from the Freesound audio tagging challenge [6].

Evolution of prediction. Figure 9 shows the evolution of predictions along the blur path on a violin audio sample. As the blur increases, prediction transforms from violin to synthesizer and then to singing bowl. The synthesizer and the singing bowl are more smooth sounding than the violin, indicating that the path is 'natural', increasing the likelihood of a good explanation.

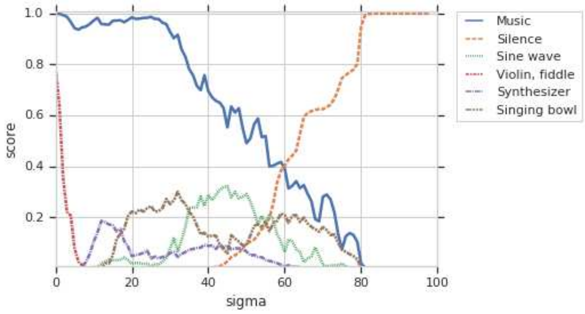


Figure 9: Visualizing the evolution of class prediction probabilities for prominent categories along the blur integration path from a violin audio sample. Y axis shows the confidence score, and X axis the sigma for the gaussian blur kernel. Color indicates the class. Initially model has higher confidence on violin (and music) class. With increased blur, confidence shifts towards synthesizer, singing bowl, sine wave, and then silence.

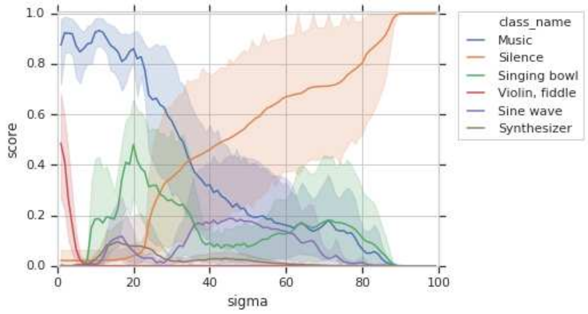


Figure 10: Visualizing the evolution of class prediction probabilities for prominent categories along the blur integration path from a few (7) violin audio samples. Y axis shows the confidence score, and X axis the sigma for the gaussian blur kernel. Color indicates the class. Initially model has higher confidence on violin and string instrument classes. With increased blur, confidence shifts towards singing bowl, sine wave, and then silence.

Class conditioning. Next, we show how Blur IG can be used to identify insights. We examine Blur IG explanations for the same audio sample. The model is a multi-label classifier and tags the piece both as 'Music' and as 'Violin'. We use Blur IG to explain both predictions. We would expect the model to predict music *because* it predicts violin, and therefore, the attributions for the classes to match. However, we find that the model looks at *different* frequencies for the two classes! Figure 11 shows the attributions. The model looks at lower frequencies for music than violin. For target class violin, the sum of the Blur IG explanations for the higher frequencies (top half across entire clip) is positive while that of the lower frequencies is negative (ratio is $\approx 19 : -1$). Whereas for class music, the sum of contributions of the higher frequencies is negative while that of the lower frequencies is positive (ratio is $\approx -33 : 1$). This is also easily observed in Figure 12 which aggregates attributions within different frequency bins (for multiple violin samples). This insight demonstrates the utility of inspecting the frequency domain.

6. Discussion

Explanations are associated with perturbations. For instance, the explanation that 'Event A caused Event B' implies that *had Event A not occurred*, Event B would not either; the phrase in italics is the perturbation associated with the explanation. The Counterfactual Theory of Causation by [13] formalizes this argument; here, the word counterfactual is synonymous with perturbation.

Feature importance/attributions are a form of explanation and indeed all the techniques that compute attributions use some type of perturbation. A good attribution technique should have two characteristics:

- For the attribution to produce a human intelligible explanation, we like the perturbation to be one that involves changing a human intelligible feature.

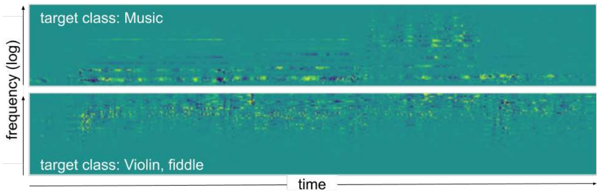
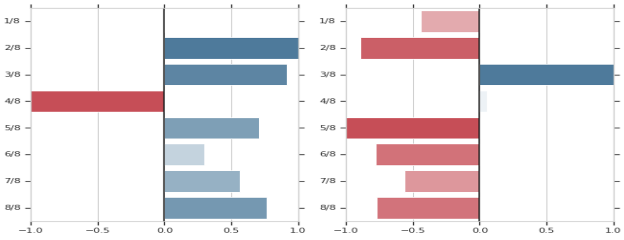


Figure 11: Blur IG explanations on a violin audio sample with target class as 'Music' (*top*) and target class as 'Violin, fiddle' (*below*). Yellow indicates positive gradients and blue the negative gradients. Explanation for class Music focuses on the lower frequencies while explanation for 'Violin, fiddle' is on the higher frequencies.



(a) Target class: 'Music'

(b) Target class: 'Violin, fiddle'

Figure 12: Aggregation of Blur IG contributions on 7 violin audio samples with target class as 'Music' (*left*) and target class as 'Violin, fiddle' (*right*). Y axis depicts the frequency bins, and X axis the integrated gradients. Blue indicates positive gradients/contributions and red the negative gradients. Explanation for class Music shows positive contributions from the lower frequencies while explanation for 'Violin, fiddle' shows positive contributions only from the higher frequencies.

- To ensure that the explanation does not produce artifacts, we would like the perturbations that only destroy information.

Blur Integrated Gradients is a *best-effort* to satisfy both conditions. Blur Integrated Gradient relies on Gaussians and Laplacians of the Gaussian to construct explanations. It is well-known that operators built using Gaussians and Laplacians of Gaussians detect edges, blogs and textures (*e.g.* [14]), *i.e.*, features that are human intelligible. Indeed, there is also biological evidence that humans rely on Gaussians and Laplacians of Gaussians to perform visual and auditory tasks [17, 18, 38].

On the computer vision front, Gaussian and Laplacians of Gaussians are workhorses of multi-scale processing. Convolutional deep neural networks used for perception tasks are built with this intuition. As the Inception paper [34] says: *'To optimize quality, the architectural decisions were based on ... and the intuition of multi-scale processing.'* By choosing an explanation method that fits the *form* of deep networks and what humans consider features, we hope to generate explanations that are simultaneously faithful to the network and intelligible. As a side-effect, we are also able to generate explanations in scale/frequency. And we show with examples in vision and audio.

References

[1] github.com/ankurtaly/integrated-gradients. 4

[2] R. J. Aumann and L. S. Shapley. *Values of Non-Atomic Games*. Princeton University Press, 1974. 2

[3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, pages 1803–1831, 2010. 1

[4] Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. *CoRR*, 2016. 1

[5] R. Fong, M. Patrick, and A. Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 2

[6] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel PW Ellis, Xavier Favory, Jordi Pons, and Xavier Serra. General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline. *arXiv preprint arXiv:1807.09902*, 2018. 7

[7] Eric J. Friedman. Paths and consistency in additive cost sharing. *International Journal of Game Theory*, 32(4):501–518, 2004. 2, 4

[8] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017. 7

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[10] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. 7

[11] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viegas, and Michael Terry. Xrai: Better attributions through regions. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 4, 6

[12] Jan J. Koenderink. The structure of images. *Biological Cybernetics*, 50(5):363–370, Aug 1984. 2, 3

[13] David K. Lewis. *Counterfactuals*. Blackwell, 1973. 8

[14] T. Lindeberg. Scale-space for discrete signals. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(3):234–254, Mar. 1990. 1, 2, 3, 4, 8

[15] Tony Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision*, 11(3):283–318, Dec 1993. 2

[16] Tony Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, Norwell, MA, USA, 1994. 1

[17] Tony Lindeberg. A computational theory of visual receptive fields. *Biological Cybernetics*, 107(6):589–635, Dec 2013. 8

[18] Tony Lindeberg and Anders Friberg. Scale-space theory for auditory signals. In Jean-François Aujol, Mila Nikolova, and Nicolas Papadakis, editors, *Scale Space and Variational Methods in Computer Vision*, pages 3–15, Cham, 2015. Springer International Publishing. 8

[19] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. 2

[20] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, 2017. 1, 2

[21] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4768–4777. Curran Associates, Inc., 2017. 1

[22] Alexander Mordvintse, Christopher Olah, and Mike Tyka. Inceptionism: Going Deeper into Neural Networks, 2015. 2

[23] Doina Precup and Yee Whye Teh, editors. *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*. PMLR, 2017. 9

[24] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. 2

[25] 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. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pages 211–252, 2015. 5

[26] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 5

[27] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Precup and Teh [23], pages 3145–3153. 1, 2

[28] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, 2013. 1, 2

[29] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. 2

[30] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, 2014. 1

[31] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Precup and Teh [23], pages 3319–3328. 1, 2, 3, 4, 5

[32] Mukund Sundararajan, Jinhua Xu, Ankur Taly, Rory Sayres, and Amir Najmi. Exploring principled visualizations for deep network attributions. 2019. 5

[33] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 4278–4284. AAAI Press, 2017. 7

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, 2014. 8

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,*, 2016. 5

[36] Gulshan V, Peng L, Coram M, and et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016. 1, 7

[37] Andrew P. Witkin. Scale-space filtering. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'83, pages 1019–1022, San Francisco, CA, USA, 1983. Morgan Kaufmann Publishers Inc. 2, 3

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 8