Visualizing and Generalizing Integrated Attributions *

Ethan Payne[1 [0000−0002−6620−8336]], David Patrick[1,2 [0000−0003−2556−8818]], and Amanda S. Fernandez[1 [0000−0003−2397−0838]]

[1] The University of Texas at San Antonio, TX, USA
[2] Texas State University, San Marcos, TX, USA

Abstract. Explainability and attribution for deep neural networks remains an open area of study due to the importance of adequately interpreting the behavior of such ubiquitous learning models. The method of expected gradients [10] reduced the baseline dependence of integrated gradients [27] and allowed for improved interpretability of attributions as representative of the broader gradient landscape, however both methods are visualized using an ambiguous transformation which obscures attribution information and neglects to distinguish between color channels. While expected gradients takes an expectation over the entire dataset, this is only one possible domain in which an explanation can be contextualized. In order to generalize the larger family of attribution methods containing integrated gradients and expected gradients, we instead frame each attribution as a volume integral over a set of interest within the input space, allowing for new levels of specificity and revealing novel sources of attribution information. Additionally, we demonstrate these new unique sources of feature attribution information using a refined visualization method which allows for both signed and unsigned attributions to be visually salient for each color channel. This new formulation provides a framework for developing and explaining a much broader family of attribution measures, and for computing attributions relevant to diverse contexts such as local and non-local neighborhoods. We evaluate our novel family of attribution measures and our improved visualization method using qualitative and quantitative approaches with the CIFAR10 and ImageNet datasets and the Quantus XAI library.

Keywords: Attribution · Saliency · Influence · Integrated Gradients · Expected Gradients · Explainability · Causal Inference · Visualization

1 Introduction

While gradient-based approaches to feature attribution for deep neural networks are both intuitive and relatively easy to implement, established methods such as

* Supported by the National Science Foundation under Grant No. 2134237. Code: https://github.com/UTSA-VAIL/Visualizing-and-Generalizing-Integrated-Attributions

integrated gradients [27] which rely on paths to fixed external reference inputs often lack a compelling justification for why certain baselines should be chosen over others. There may be situations and applications which may support obvious baselines, but as noted by Erion et al. [10], this is often not the case. Many of the shortcomings of integrated gradients were alleviated by computing the expected gradients as a Monte Carlo integral over the training dataset, however this approach does not succeed in completely generalizing the original intuition of integrated gradients to a comprehensive family of attribution measures.

We present a generalization of integrated gradients [27] and expected gradients [10] which also encompasses a diverse family of other attribution measures. By formulating the expected gradients in terms of a volume integral rather than a path integral, we obtain an attribution method which is immediately generalizable to any deep learning application, and which can be easily iterated upon. We note that our formulation has similar implementation requirements as expected gradients while allowing us to access several unique sources of attribution information which were previously not utilized. Using our new formulation of *generalized integrated gradients*, we are able to identify distinct paradigms of attribution information corresponding to input locality.

Additionally, leverage our new formulation to develop three new measures of gradient variance, stability, and consistency, which each quantify a unique aspect of model behavior. Gradient variance quantifies the dispersion of model gradients, and results in attributions which provide improved visual salience over expected gradients. Our stability and consistency measures incorporate angular information to characterize the behavior of model gradients, with stability quantifying whether the input is a local optimum, and consistency quantifying disagreement between gradients at different locations in the space.

Finally, considering that the interpretation of image attributions depends heavily on their semantic interpretation, we propose a new procedure for visualizing attributions which addresses several concerns associated with the visualization methods commonly employed in the past. Notably, we address the problems of artificial introduction of information from reference inputs, loss of color channel-specific information, and loss of attribution sign.

Using our new visualization procedure, we present our proposed measures qualitatively evaluated on ImageNet [8] using gradients from a pre-trained ResNet-34 model, as illustrated in Figure 1 with evaluation examples shown in Figure 2. In summary, our contributions include:

- A method of more accurately and faithfully visualizing attributions
- A mathematical formulation to describe and develop a generalized family of novel integrated attribution measures
- Several specific useful measures of interest constructed from descriptive statistics using our formulation

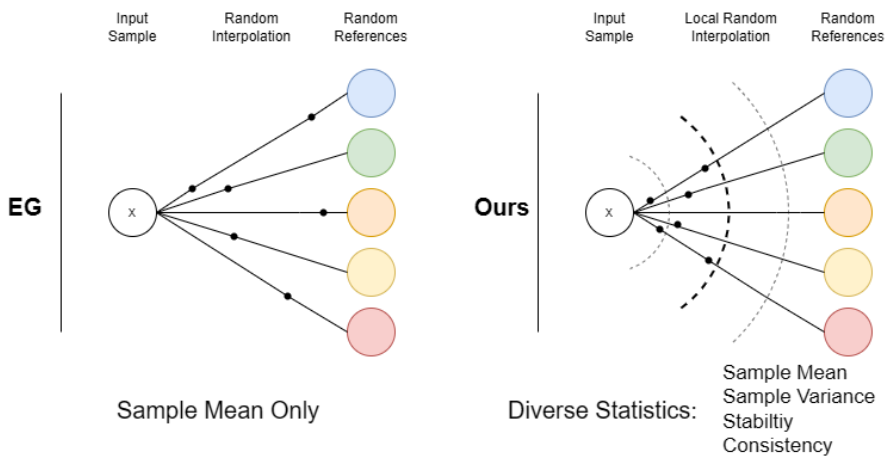


Fig. 1: Illustration of our method as compared with Expected Gradients [10]. We notably include a locality parameter as well as the ability to compute additional descriptive statistics beyond a simple sample mean.

2 Related Work

Early methods in explainability, such as layer-wise relevance propagation (LRP) [3], decompose the predictions of nonlinear classifiers to obtain attributions for individual pixels. Many current methods utilize various forms of gradient information in order to generate attributions [2,1]. In an effort to increase the robustness of these feature attributions, Sundararajan et al. [27] selected a set of axioms to guide the development of a more robust attribution measure which they call integrated gradients. Integrated gradients are computed by taking a linear path from an input of interest to a baseline input, and integrating the gradients of the model with respect to the input over this path, as is discussed in greater detail below in Section 3.1. To allow for efficient computation of integrated gradients, Hesse et al. [14] consider a special class of nonnegatively homogenous deep neural networks, and to remove the arbitrary baseline selection issues associated with integrated gradients. With their 'iterated integrated attributions [5], Barken et al. utilize linear interpolations of the input as well as intermediate representations from within the model. Erion et al. [10] use examples from the training dataset as baselines, which re-contextualizes the resulting attribution values as the expectation of model gradients over the data, with similar approach being taken by Lundberg et al. [19] to approximate Aumann-Shapley (SHAP) values. Merrill et al. define a "generalized integrated gradients" [21] from an axiomatic, algebraic perspective in the context of Aumann-Shapley values in order to extend the concept of path-integrated credit assignment to more diverse function spaces such as those relevant to applications in finance. While we also define a "generalized integrated gradients" in this work, ours is instead framed in the context of developing a broader family of integrated attribution measures of which

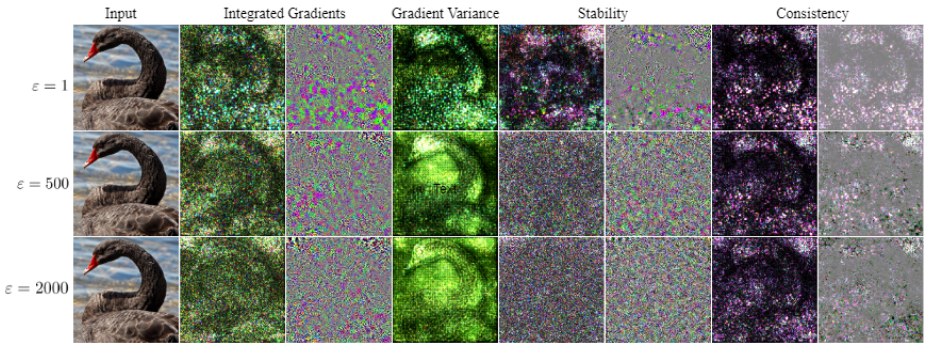


Fig. 2: Summary of our newly proposed family of attribution measures and visualization methods [best viewed in color]. Each measure is computed using the locality method of Equation 4, and the sampling method of expected gradients [10] using 500 sample points (see Figure 4 for additional sampling details). From top to bottom: $\varepsilon = 1, \varepsilon = 500, \varepsilon = 2000$. From left to right: input, local integrated gradients (unsigned), local integrated gradients (signed), gradient variance (unsigned only), stability (unsigned), stability (signed), consistency (unsigned), consistency (signed).

the path-integrated gradients is a special case. Extending prior work in attribution to include hidden units within a neural network, Dhamdhere et al. [9] introduce the notion of *conductance*. This neuron attribution builds on the integrated gradients attribution method, with conductance being formulated as the flow of integrated gradients via a given hidden unit. This work on neuron conductance is refined by Shrikumar et al. [26], who develop a scalable implementation they call neuron integrated gradients. In another instance of attribution methods being applied towards other deep learning tasks, Jha et al. [15] construct an attribution-based confidence (ABC) metric for measuring whether an output can be trusted. Variants of the metric utilize different attribution methods, one being integrated gradients. Hase et al. [12] also compare several salience-based explanation methods (such as integrated gradients) and several search-based methods such as their parallel local search. In particular, they posit that the use of out-of-distribution counterfactual inputs like the baselines required for integrated gradients is problematic. Our proposed generalized method builds on the success of expected gradients [10] in addressing the above concerns regarding the out-of-distribution counterfactual inputs which are often used in attribution methods, and enables further development of nuanced attribution measures.

3 Generalized Integrated Attributions

Visualization of Pixel Attributions We first discuss the approach we have taken for visualizing pixel attributions for computer vision tasks, as this has been an area of significant recent interest [1,24] and is essential for the accurate

interpretation of computed attributions. Any transformation of attribution values which is not invertible will result in loss of information by compression, as will any transformation which introduces information from an outside source in the form of noise. Previous methods, such as integrated gradients [27] and expected gradients [10], chose to visualize computed attribution values by taking the absolute value (compression), aggregating values for each color channel to a single per-pixel attribution (compression), clipping extreme values (compression), scaling to the range $[0, 1]$, and then multiplying the resulting values by the original input image (noise). Perhaps most importantly, multiplication by the input results in an extremely misleading attribution visualization which artificially resembles the original input image (see Figure 3). Furthermore, the choice of aggregating color channels needlessly obscures channel-dependent information, which demonstrate to be highly informative. While clipping to quantiles and rescaling to a given range may often be necessary to produce visualizations perceptible to human users, we should always make careful note of these transformations and remind ourselves that each of these transformations may reveal or obfuscate unique sources of information.

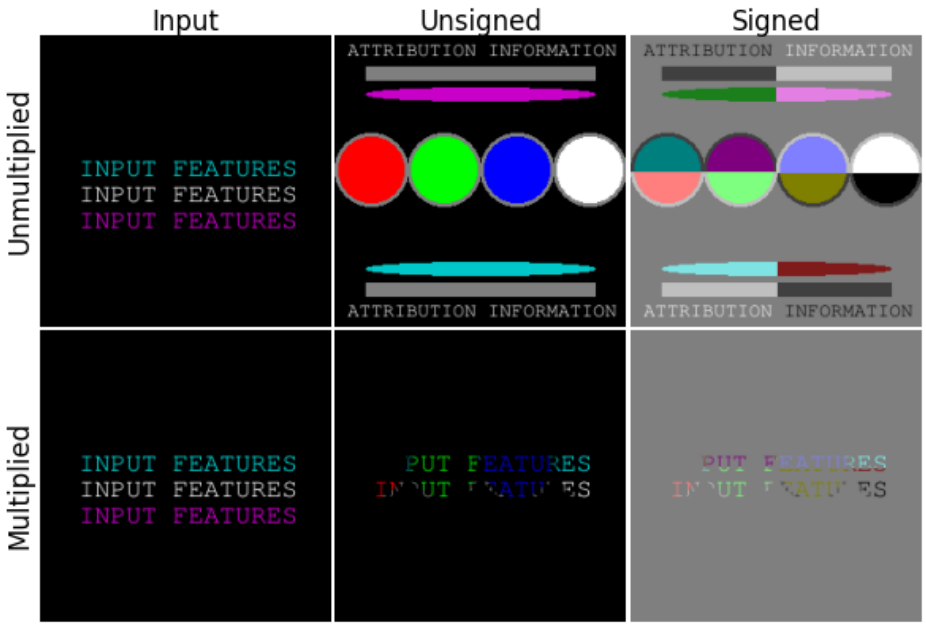


Fig. 3: Comparison of visualization methods [best viewed in color]. We consider a hypothetical input (column 1) and a hypothetical attribution consisting of a test pattern with both positive and negative values to illustrate the difference between signed and unsigned approaches. We can observe that multiplying by the input results in a significant loss of information and bias towards the input.

Unsigned Visualization When we are interested in the magnitude of attribution values and not whether they are positive or negative, we can take the absolute value of the attributions and scale them to $[0, 1]$ after first clipping extreme values. This preserves color channel information and introduces no artificial information from the original input. Using this method, attributions with small magnitude are dark while attributions with large magnitude are bright (see row 1, column 2 in Figure 3).

Signed Visualization In contrast to unsigned visualization, if we wish to visualize the difference between positive and negative attribution values, we instead scale the attributions to $[-1, 1]$ after clipping extreme values. Then, we selectively brighten or darken a blank slate image starting from 50% uniform brightness to obtain the final attribution map. This method preserves both the sign of the attributions and all color-dependent information while introducing no artificial bias from the original input. Using this method, negative attributions are dark while positive attributions are bright (see row 1, column 3 in Figure 3).

As demonstrated in Figure 3, there are unique advantages and disadvantages to both signed and unsigned attribution visualization, and ideally both should be used in concert when interpreting attribution results. Importantly, any visualization of attribution measures should not be obscured by any information from a particular reference input unless absolutely necessary, in the interest of introducing as little bias as possible into the final interpretation of a given attribution result. In cases where an unambiguous mask can be constructed from prediction attributions, such a mask might be used to highlight regions of a particular reference input, but this masking should be performed with caution and careful consideration in order to avoid the misinterpretation of input features as attribution results.

3.1 Extending Expected Gradients

The reformulation of integrated gradients as an expected value developed by Erion et al. [10] allows the original path integrals of Sundararajan et al. [27] to be completely discarded in favor of volume integrals over the input space. However, this simplification was not thoroughly realized in the presentation of expected gradients. We now reformulate integrated gradients as a generalized integral over a volume in the input space. Sundararajan et al. [27] defines the *path integrated gradients* (Equation 1) for a model F and path function $\gamma(\alpha), \quad \alpha \in [0, 1]$ from the input x_0 to a baseline which we recall below:

$$\text{PathIntegratedGrads}_{\gamma}(x_0) ::= \int_{\alpha=0}^{1} \frac{\partial F(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \frac{\partial \gamma_i(\alpha)}{\partial \alpha} d\alpha \tag{1}$$

Erion et al. [10] extends this with the method of *expected gradients*, which aggregates the path integrated gradients for a distribution of many paths γ, and specifically considers a collection of paths using a uniform distribution over examples from the training set as baseline path endpoints. We now define the

generalized integrated gradients (Equation 2) over a set $\mathbb{S}$ and a probability density function $p_{\mathbb{S}}$:

$$\begin{aligned} \text{GeneralizedIntegratedGrads}(\mathbb{S}) ::&= \mathbb{E}_{\mathbb{S}}\left[\nabla F\right] \\ &= \int_{\mathbb{S}} \nabla F(x) p_{\mathbb{S}}(x) dx \end{aligned} \tag{2}$$

If we follow the method of expected gradients [10] and assume a uniform distribution over $\mathbb{S}$ with $|\mathbb{S}|$ the volume (or even more generally the Lebesgue measure) of $\mathbb{S}$, we obtain Equation 3:

$$\text{GeneralizedIntegratedGrads}(\mathbb{S}) ::= \frac{1}{|\mathbb{S}|} \int_{\mathbb{S}} \nabla F(x) dx \tag{3}$$

The generalized formulation of Equation 2 includes the expected gradients [10] as a special case, which in turn includes the path-based integrated gradients [27] as a special case. To illustrate an immediate advantage over expected gradients, we define below the *local integrated gradients* (Equation 4) for a neighborhood $\mathcal{B}_{\varepsilon}(x_0)$, i.e. the n-dimensional ball of radius ε centered on an input x_0, where n is the number of dimensions of the input, and $V_n(\varepsilon)$ is the volume of the n-dimensional ball of radius ε. Notice that for $\varepsilon = \infty$, this method is equivalent to expected gradients when the space is sampled along paths γ between the input x and examples from the training dataset, but other volume sampling methods are now available for exploration. Importantly, by controlling the radius ε, we are now also able to collect the integrated gradients corresponding to a specific locality (Figures 4, and 8a), and we can do the same for the other descriptive statistics which we develop below (Figures 5, 8b, 6, 8c, 7, 8d).

$$\text{LocalIntegratedGrads}(x_0, \varepsilon) ::= \frac{1}{V_n(\varepsilon)} \int_{\mathcal{B}_{\varepsilon}(x_0)} \nabla F(x) dx \tag{4}$$

We then to compute a numerical approximation of the desired integral over the desired set. We can again follow the example of expected gradients [10] and collect sample points S within the set $\mathbb{S}$ to approximate with a Monte Carlo integral as follows in Equation 5, where $|\mathbb{S}|$ is the volume of the set $\mathbb{S}$, and $|S|$ is the number of points in the sample S.

$$\begin{aligned} &\text{GeneralizedIntegratedGrads}(\mathbb{S}) ::= \frac{1}{|\mathbb{S}|} \int_{\mathbb{S}} \nabla F(x) dx \\ &\sim \frac{1}{|\mathbb{S}|} \left[\frac{|\mathbb{S}|}{|S|} \sum_{s \in S} \nabla F(s) \right] = \frac{1}{|S|} \sum_{s \in S} \nabla F(s) = \mathbb{E}_S\left[\nabla F\right] \end{aligned} \tag{5}$$

3.2 Novel Attribution Measures

Using the above framework developed for generalizing integrated gradients (Equation 2), we now propose three new feature attribution measures as descriptive statistics which account for different aspects of model behavior. Again assuming

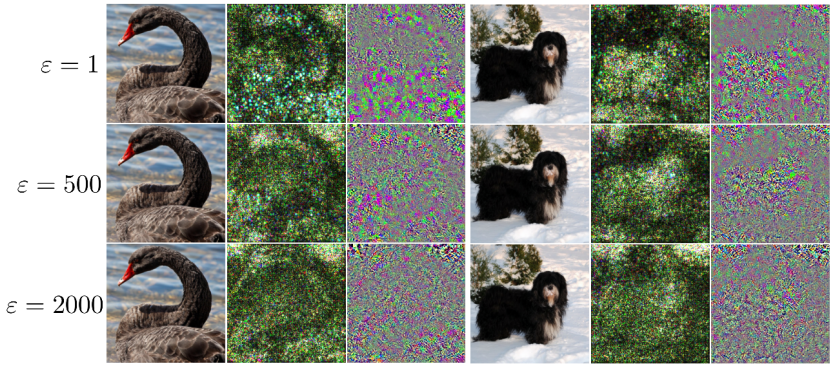


Fig. 4: Local integrated gradients of Equation 4 [best viewed in color]. We can observe how the choice of ε results in noticeably different attributions, and how the unsigned and signed visualizations reveal different patterns especially with respect to color channels. We compute this measure for $\varepsilon = 1, \varepsilon = 500, \varepsilon = 2000$ (top, middle, bottom row respectively). Immediately to the right of the input are the attributions visualized using our unsigned method. We sample $\mathcal{B}_\varepsilon(x_0)$ using a reference dataset as in the method of expected gradients [10], using 100 reference elements and 5 uniform random sample points on each of these vectors within the ball $\mathcal{B}_\varepsilon(x_0)$, for a total of 500 sample points, yielding a *local expected gradients*.

a uniform distribution over $\mathbb{S}$, Monte Carlo approximation with a sample set S can be applied for each of these measures as easily as for generalized integrated gradients by following the example of Equation 5. If we follow the method of selecting $\mathbb{S}$ used for local integrated gradients 4, we can also again compute all of the following measures according to a desired locality radius ε.

Gradient Variance Building on the formulation of integrated gradients as a sample mean by Erion et al. [10], we now construct a sample variance (Equation 6) to quantify the dispersion of model gradients over the set $\mathbb{S}$. Note that we again are able to preserve color channel information, but since variances are strictly positive measures, we do not need to consider visualizing negative values (Figures 5 and 8b).

$$\begin{aligned} &\text{GradientVariance}(\mathbb{S}) \\ &\quad ::= \frac{1}{|\mathbb{S}|} \int_{\mathbb{S}} \left(\nabla F(x) - \mathbb{E}_{\mathbb{S}}\left[\nabla F(x)\right]\right)^2 dx \\ &\quad = \mathbb{E}_{\mathbb{S}}\left[\nabla F(x)^2\right] - \mathbb{E}_{\mathbb{S}}\left[\nabla F(x)\right]^2 \end{aligned} \tag{6}$$

Stability We propose a measure of local stability as follows (Equation 7). For each sample point s within the set $\mathbb{S}$, we compute the vector $s - x_0$ defining

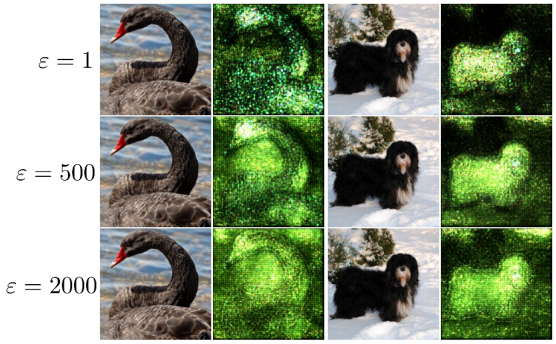


Fig. 5: Gradient variance of Equation 6 [best viewed in color]. We can again observe the effect of the locality radius ε and the presence of color-dependent patterns. We also obtain visualization which are significantly more salient than those we obtained for local expected gradients (Figure 4). We use use the same sample scheme and choices of ε as in Figure 4. Since variances are strictly positive, we only use our unsigned visualization method (Section 3).

the offset of this sample point from the original input. We then compute the cosine similarity of between the offset vector and the gradients at the sample point $\nabla F(s)$. The intuition of this measure is that if the gradients at a sample location point back toward the input, then that input can be considered 'stable', in that the input is a local optimum. The total stability measure is taken as the expectation of these angles over the set $\mathbb{S}$ as:

$$\begin{aligned} \text{Stability}(\mathbb{S}, x_0) ::= & \frac{1}{|\mathbb{S}|} \int_{\mathbb{S}} \frac{(x - x_0) \cdot \nabla F(x)}{\|(x - x_0)\| \|\nabla F(x)\|} dx \\ & = \mathbb{E}_{\mathbb{S}} \left[\cos(\theta)\right], \\ & \theta \text{ the angle between } \nabla F(x) \text{ and } (x - x_0) \end{aligned} \tag{7}$$

To avoid losing channel-dependent information, we compute three angles $(\theta_{rg}, \theta_{gb}, \theta_{br})$ using pairs of pixels as 2-dimensional vectors. We map the values θ_{rg} to the blue channel, θ_{gb} to the red channel, and θ_{br} to the green channel for Figure 6.

Consistency Finally, we propose a measure which we call 'consistency' (Equation 8). For each sample point s within the set $\mathbb{S}$, we compute the cosine similarity of the gradients of the model at the sample point $\nabla F(s)$ and the gradients at the input $\nabla F(x_0)$. The intuition of this measure is that if the gradients at a sample location point in the same direction as the gradients at the input, then the model gradients are locally consistent with each other. The total consistency measure is taken as the expectation of these angles over the set $\mathbb{S}$ as:

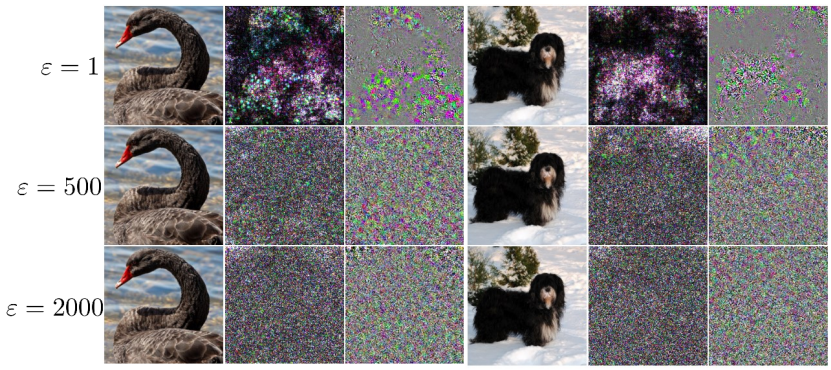


Fig. 6: Stability measure of Equation 7 [best viewed in color]. We only observe salient images for small ε, as for larger ε the input x_0 is likely no longer a local optimum. We use use the same sample scheme and choices of ε as Figure 4.

$$\begin{aligned} \text{Consistency}(\mathbb{S}, x_0) :: &= \frac{1}{|\mathbb{S}|} \int_{\mathbb{S}} \frac{\nabla F(x) \cdot \nabla F(x_0)}{\|\nabla F(x)\| \|\nabla F(x_0)\|} dx \\ &= \mathbb{E}_{\mathbb{S}} \left[\cos(\theta)\right], \\ \theta \text{ the angle between } &\nabla F(x) \text{ and } \nabla F(x_0) \end{aligned} \tag{8}$$

Again, we preserve the color-dependent information by computing three angles $(\theta_{rb}, \theta_{rg}, \theta_{bg})$ using pairs of pixels as 2-dimensional vectors, and mapping the similarity value representing a given pair of channels to the remaining channel for the final visualization (Figure 7).

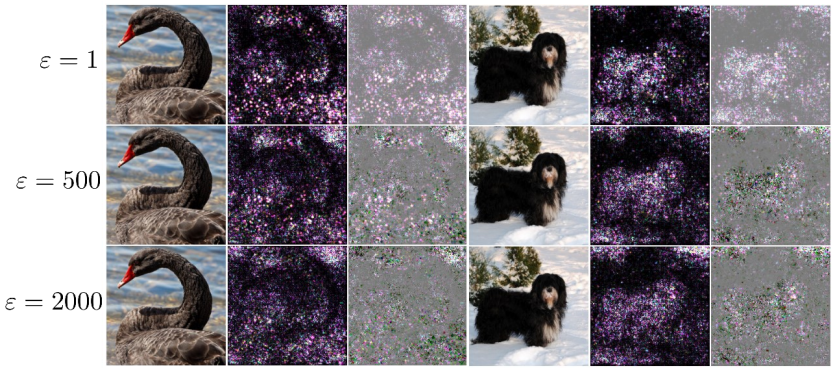
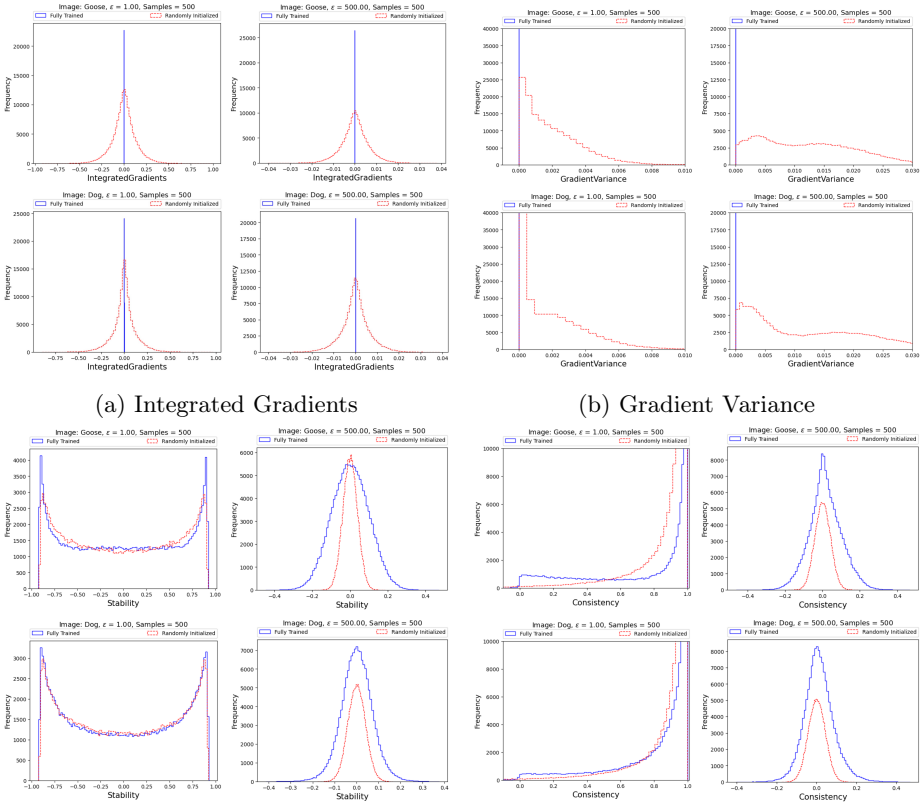


Fig. 7: Consistency measure of Equation 8 [best viewed in color]. This measure allows for determining which pixels the gradients at nearby images $x \in \mathcal{B}_\varepsilon(x_0)$ either agree or disagree with the gradients at the image x_0. The same sample scheme and choices of ε are the same as in Figure 4.



(a) Integrated Gradients (b) Gradient Variance

(c) Stability (d) Consistency

Fig. 8: Histograms of each of our novel measures corresponding to gradients from both a randomly initialized and fully-trained ResNet-34 (Row 1: goose, Row 2: dog). We can observe some recognizable parametric families and different paradigms for small and large ε, with a clear distinction between the trained (blue) and untrained (red) models. If attribution values converge in distribution during model training, this may reveal valuable insight regarding future training optimizations, heuristics, and diagnostics.

Generalized Integrated Attributions In the interest of describing all of the above measures as well as any similarly constructed descriptive statistic using a single unified formulation, we provide the following definition of a *generalized integrated attribution* (Equation 9). By selecting an attribution function $\mathcal{A}$, a model F, a set of interest $\mathbb{S}$, and a probability density function $p_{\mathbb{S}}$, we can access a limitless number of unique statistics to describe high-dimensional gradient landscapes.

$$\begin{aligned} &\text{GeneralizedIntegratedAttribution}(\mathcal{A}, F, \mathbb{S}, p_{\mathbb{S}}) \\ &::= \int_{\mathbb{S}} \mathcal{A}(F, x) p_{\mathbb{S}}(x) dx \end{aligned} \tag{9}$$

Note that we do not necessarily include a particular input x_0 as a required argument, as we can in theory compute attributions over entire sets $\mathbb{S}$ without referring directly to any single input. For the case of local integrated gradients, the set of interest $\mathbb{S}$ is the ε-ball centered at an input x_0, but this is a justification for the choice of $\mathbb{S}$. Note that our stability and consistency measures appear to require an input x_0, but these can be framed instead as particular choices of attribution function $\mathcal{A}$.

While many interesting attribution measures such as the several new measures we have introduced above are described by the family of generalized integrated attributions, there are likely many more complex attributions of interest which cannot be formulated concisely as a single integral or expected value. Nevertheless, this new formulation can assist in the classification and analysis of newly-developed attribution measures.

4 Evaluation Using Quantus [13]

In addition to providing the above qualitative attribution outputs, we also consider a quantitative evaluation of our approach, although there is still no broad consensus regarding reliable metrics for attribution [1,24,16]. We provide some quantitative results in Table 1 using the Quantus XAI library, which provides a toolkit of various attribution methods and evaluation. Metrics in this library are organized into several broad categories such as Faithfulness, Robustness, and Complexity. Given that each metric is unique and sensitive to its own hyperparameters, detailed descriptions defining each method are provided by Hedstrom et al. [13]. We evaluated each attribution method on the full CIFAR-10 [18] test set, using a pre-trained ResNet-18 model.

5 Conclusion

In this work, we present a generalized formulation of the feature attribution methods integrated gradients and expected gradients by contextualizing expected values as general integrals over sets of interest. Furthermore, we demonstrate how this approach makes available new sources of attribution information, such as differences between local and nonlocal attribution paradigms, and novel attribution measures. This framework also allows for new forms of parametric control over attribution measures such as the choice of locality radius ε and the sampling distribution over the set $\mathbb{S}$. Overall, this new formulation of integrated attributions represents a significant transition towards a much broader family of generalizable measures. Additionally, we introduce a novel method for visualizing attributions which addresses information loss in current approaches. Such

Method	ε	Faithfulness (↑)		Robustness (↓)		Complexity	
		PixFlip	FaithCorr	MaxSens	AvgSens	Sparse(↑)	Complex(↓)
Integrated Gradients [27]	n/a	0.23133	0.04774	0.13018	0.11247	**0.59017**	**6.29801**
Saliency [23,4]		0.28260	0.03239	0.13332	0.11957	0.43868	6.60204
GradientShap [20]		0.23266	0.04752	0.18278	0.14631	0.58966	6.29854
FeatureAblation [17]		0.18525	0.13089	0.11974	0.10510	0.58176	6.32653
FeaturePermutation [11]		0.16536	**0.14338**	0.19927	0.18554	0.55717	6.38713
Deconvolution [29]		0.30896	-0.00627	**1.9e-08**	**1.8e-08**	0.51399	6.48971
Expected Gradients	1	0.24490	0.02238	1.13295	1.03500	0.50759	7.58803
	10^3	0.23667	0.01751	1.33943	1.07549	0.46421	7.66907
Gradient Variance	1	**0.34443**	0.04312	0.78553	0.65590	0.56980	7.41242
	10^3	0.27669	0.03585	1.12104	0.80185	0.46126	7.65275
Stability	1	0.28154	0.01003	1.42457	1.25999	0.41586	7.74354
	10^3	0.28020	-0.00411	1.02010	0.99323	0.41840	7.74035
Consistency	1	0.27990	-0.00454	0.30972	0.29827	0.10239	7.99951
	10^3	0.28233	-0.00418	1.16962	1.09159	0.39830	7.76662

Table 1: Quantitative evaluation of novel attribution measure family using the Quantus XAI library [13]. Metrics used are: PixelFlipping [3], FaithfulnessCorrelation [6], MaxSensitivity [28], AvgSensitivity [28], Sparseness [7], Complexity [6]. Results are averaged over the CIFAR10 [18] test set. Our (local) Expected Gradients, Gradient Variance, Stabiltiy, and Consistency measures were each computed by Monte Carlo integration using 100 sample points within the ball of radius ε.

approaches to more explainable AI can have significant societal impact, enabling better transparency and bias mitigation than treating learning models as black boxes. Our work to reduce misinformation and bias in feature attributions directly addresses the growing need for transparency and fairness with respect to machine learning.

Limitations Our method depends heavily on Monte Carlo integration, therefore the accuracy, computational efficiency, and robustness of our attribution results likewise depend on the design and incorporation of effective numerical integration schemes. Specifically, for large sets $\mathbb{S}$, or equivalently large radius ε, the number of sample points required to obtain a good approximation of the true integral increases exponentially. Similarly, any axiomatic properties of our family of measures would also depend on a good approximation of the underlying integral, so this poses a computational challenge to scaling if we desire to measure attributions over large sets. Note however, that other state-of-the-art methods such as expected gradients methods have similar numerical scaling limitations.

Future Work Numerical techniques, such as those developed by Mitchell et al. [22], Reeger et al. [25], and Hesse et al. [14], may serve to improve the efficiency and accuracy of integrated attributions. Additionally, we can conduct convergence analyses for hyperparameters such as the sample size and the locality radius ε, and we can explore the metrics based on Aumann-Shapley values developed by Lundberg et al. [19]. In addition, we should assess our new family

of measures using an analytic or algebraic approach similar to the selection of desirable axioms by Sundararajan et al. [27] and Merrill et al. [21]. Erion et al. [10] made another significant contribution with their method of using attribution prior for training regularization, so we should apply this technique to train models using our new measures for these attribution priors. To explore additional sources of model attribution, and since integrated gradients forms the basis for layer conductance [26], we should develop implementations of our new measures which can be applied within the space of convolutional filters. Extending attribution measures to applicability in the abstract feature space may also have the benefit of revealing new sources of relevant attribution information.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant No. 2134237. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)
2. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for deep neural networks. In: International Conference on Learning Representations (2018)
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
4. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.R.: How to explain individual classification decisions. The Journal of Machine Learning Research **11**, 1803–1831 (2010)
5. Barkan, O., Elisha, Asher, Y., Eshel, A., Koenigstein, N.: Visual explanations via iterated integrated attributions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2073–2084 (October 2023)
6. Bhatt, U., Weller, A., Moura, J.M.: Evaluating and aggregating feature-based model explanations. In: Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence. pp. 3016–3022 (2021)
7. Chalasani, P., Chen, J., Chowdhury, A.R., Wu, X., Jha, S.: Concise explanations of neural networks using adversarial training. In: International Conference on Machine Learning. pp. 1383–1391. PMLR (2020)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Dhamdhere, K., Sundararajan, M., Yan, Q.: How important is a neuron. In: International Conference on Learning Representations (2019)
10. Erion, G., Janizek, J., Sturmfels, P., Lundberg, S., Lee, S.I.: Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nature Machine Intelligence **3**, 1–12 (2021)

11. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. **20**(177), 1–81 (2019)
12. Hase, P., Xie, H., Bansal, M.: The out-of-distribution problem in explainability and search methods for feature importance explanations. Advances in Neural Information Processing Systems **34** (2021)
13. Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.C.: Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. Journal of Machine Learning Research **24**(34), 1–11 (2023)
14. Hesse, R., Schaub-Meyer, S., Roth, S.: Fast axiomatic attribution for neural networks. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 19513–19524. Curran Associates, Inc. (2021)
15. Jha, S., Raj, S., Fernandes, S., Jha, S.K., Jha, S., Jalaian, B., Verma, G., Swami, A.: Attribution-based confidence metric for deep neural networks. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
16. Jyoti, A., Ganesh, K.B., Gayala, M., Tunuguntla, N.L., Kamath, S., Balasubramanian, V.N.: On the robustness of explanations of deep neural network models: A survey. ArXiv **abs/2211.04780** (2022)
17. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., et al.: Captum: A unified and generic model interpretability library for pytorch. arXiv preprint:2009.07896 (2020)
18. Krizhevsky, A.: Learning multiple layers of features from tiny images (2009)
19. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence **2**(1), 56–67 (Jan 2020). https://doi.org/10.1038/s42256-019-0138-9
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4768–4777. NIPS'17, Curran Associates Inc. (2017)
21. Merrill, J., Ward, G., Kamkar, S., Budzik, J., Merrill, D.: Generalized integrated gradients: A practical method for explaining diverse ensembles. arXiv preprint arXiv:1909.01869 (2019)
22. Mitchell, S.A., Awad, M.A., Ebeida, M.S., Swiler, L.P.: Fast approximate union volume in high dimensions with line samples. Tech. rep., Sandia National Lab.(SNL-NM), Albuquerque, NM (United States) (2018)
23. Morch, N., Kjems, U., Hansen, L., Svarer, C., Law, I., Lautrup, B., Strother, S., Rehm, K.: Visualization of neural networks using saliency maps. In: Proceedings of ICNN'95 - International Conference on Neural Networks. vol. 4, pp. 2085–2090 vol.4 (1995). https://doi.org/10.1109/ICNN.1995.488997
24. Rao, S., Böhle, M., Schiele, B.: Towards better understanding attribution methods. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10223–10232 (June 2022)
25. Reeger, J.: Approximate integrals over the volume of the ball. Journal of Scientific Computing **83** (05 2020). https://doi.org/10.1007/s10915-020-01231-y
26. Shrikumar, A., Su, J., Kundaje, A.: Computationally efficient measures of internal neuron importance. CoRR **abs/1807.09946** (2018)

27. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
28. Yeh, C.K., Hsieh, C.Y., Suggala, A., Inouye, D.I., Ravikumar, P.K.: On the (in) fidelity and sensitivity of explanations. Advances in Neural Information Processing Systems **32** (2019)
29. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 818–833. Springer (2014)