On Smoother Attributions using Neural Stochastic Differential Equations

Sumit Jha[1] , **Rickard Ewetz**[2] , **Alvaro Velasquez**[3] and **Susmit Jha**[4]
[1]University of Texas at San Antonio, San Antonio, TX
[2]University of Central Florida, Orlando, FL
[3]Air Force Research Laboratory, Rome, NY
[4]SRI International, Menlo Park, CA
sumit.jha@utsa.edu, rickard.ewetz@ucf.edu, alvaro.velasquez.1@us.af.mil, susmit.jha@sri.com

Abstract

Several methods have recently been developed for computing attributions of a neural network's prediction over the input features. However, these existing approaches for computing attributions are noisy and not robust to small perturbations of the input. This paper uses the recently identified connection between dynamical systems and residual neural networks to show that the attributions computed over neural stochastic differential equations (SDEs) are less noisy, visually sharper, and quantitatively more robust. Using dynamical systems theory, we theoretically analyze the robustness of these attributions. We also experimentally demonstrate the efficacy of our approach in providing smoother, visually sharper and quantitatively robust attributions by computing attributions for ImageNet images using ResNet-50, WideResNet-101 models and ResNeXt-101 models.

1 Introduction

Deep neural networks (DNNs) are increasingly deployed in healthcare, security, autonomous driving, and other safety-critical applications. The responsible use of deep learning in such high-assurance applications necessitates the capability to explain their decisions. However, while a plethora of attribution and explanation techniques have been developed [Simonyan *et al.*, 2013; Sundararajan *et al.*, 2017; Shrikumar *et al.*, 2017; Lundberg and Lee, 2017; Kim *et al.*, 2018], these techniques are known to be noisy and not robust to small input perturbations [Ghorbani *et al.*, 2019]. The noise in pixel-level attributions (Figure 1, center) produced by path-integral-based methods [Sturmfels *et al.*, 2020; Smilkov *et al.*, 2017; Xu *et al.*, 2020; Dombrowski *et al.*, 2019] has recently been analyzed, and factors such as high manifold curvature and choice of baselines have been identified.

In this paper, we focus on attributions computed over deep learning models that can be analyzed using the theory of dynamical systems [Chen *et al.*, 2015; Lu *et al.*, 2018]. Residual neural networks have been modeled using neural ordinary differential equations (ODEs) [Chen *et al.*, 2018], and stochastic variants of ResNets are described using neural stochastic differential equations (SDEs) [Wang *et al.*, 2019; Liu *et al.*, 2020; Wang *et al.*, 2019]. We leverage methods from the study of stochastic dynamical systems to show theoretically and empirically that a suitable injection of noise into the residual layers during training produces robust attributions. While addition of noise in the inputs during inference has empirically been shown to improve attributions [Smilkov *et al.*, 2017], we use the theory of stochastic differential equations to inject noise at different residual layers and improve the robustness of attributions computed by path-integrals [Lundberg and Lee, 2017].

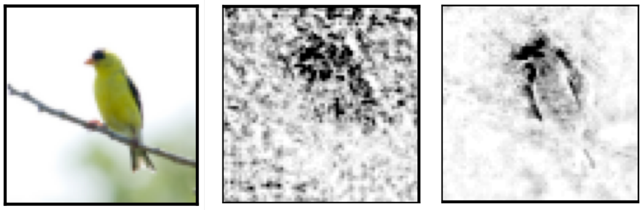

Figure 1: Input image (left) and results of DeepLIFT SHAP applied to Neural ODEs (center) and Neural SDEs (right). Neural SDEs (right) have visually sharper and less noisy (smoother) attributions.

We make the following contributions towards computing robust attributions of deep neural network decisions:

1. We perform a worst-case analysis to show that the total attribution computed on neural SDEs trained using noise are more robust than attributions computed on neural ODEs. We are the first to use the connection between dynamical systems and residual neural networks to study the quality of attributions.
2. This robustness of the attributions of our SDE approach is agnostic to the choice of path-integral attribution methods (see tables 1, 2 and 3). We quantitatively demonstrate improvement in the computed attributions using the sensitivity metric [Yeh *et al.*, 2019].
3. We experimentally illustrate that attributions computed using neural SDEs are qualitatively better, smoother and visually sharper than those obtained from neural ODEs. This improvement is demonstrated on several attribution methods including DeepLIFT SHAP, Smooth-Grad, and Integrated Gradients (see figures 1 through 4).

2 Overview

We present a brief summary of our observations and sketch our theoretical analysis as well as our experimental results.

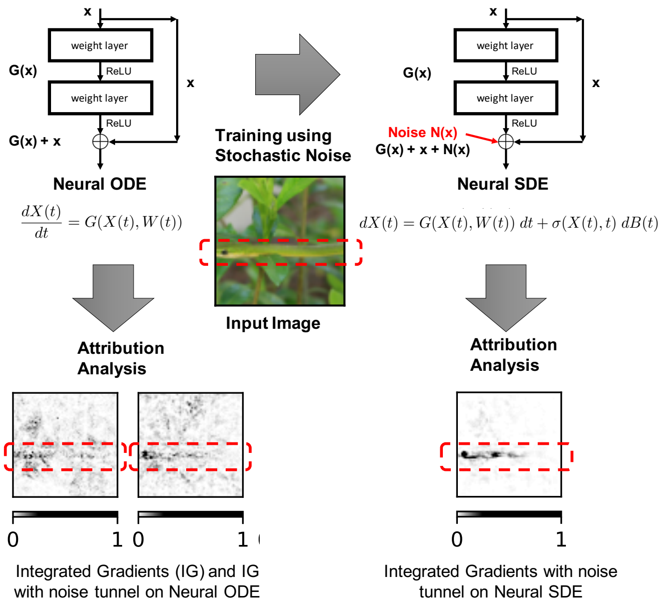

Figure 2: An overview of our observations and results: ResNets with stochastic noise injected into the residual layers or neural SDEs create more robust attributions. We show that the sum of change in attributions is smaller for neural SDEs than that for neural ODEs. The input image of a green snake on a green background and its attributions are shown; the integrated gradient attribution with noise tunnel for our neural SDE approach (bottom right) is visually sharper than the integrated gradient (IG) attribution as well as IG coupled to a noise tunnel for the neural ODE (bottom left).

Given a traditional residual network

$$X(i+1) = X(i) + R(X(i), W(i)),$$

the layers i of the network can be viewed as a time discretization of a continuous neural ODE

$$\frac{dX(t)}{dt} = G(X(t), W(t))$$

We suggest the use of neural SDEs

$$dX(t) = G(X(t), W(t))\, dt + \sigma(X(t), t)\, dB(t)$$

or ResNets with noise $N(i)$ injected at each residual layer

$$X(i+1) = X(i) + R(X(i), W(i)) + N(i)$$

This injection of noise in each residual layer of a residual network or a neural SDE leads to visually sharper and more robust attributions than the traditional residual networks or neural ODEs. Experimentally, we find that attribution methods create visually sharper attributions for neural SDEs. For example, Figure 2 shows how integrated gradients for neural SDE clearly identifies the green snake in the bottom right figure while both integrated gradients and integrated gradients with noise tunnel [Kokhlikyan *et al.*, 2020] produce diffused attributions in the bottom left figures. The snake in the input image and in the attributions is highlighted using a dashed red box.

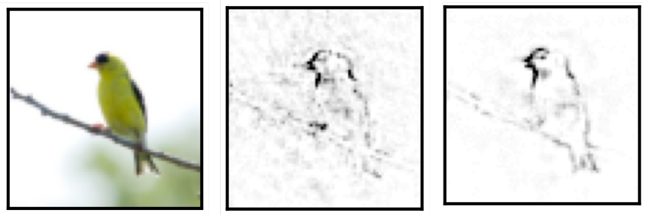

Figure 3: Input image (left) and results for Integrated Gradients with noise tunnel on neural ODE (center) and the same attribution method applied to our neural SDE (right). The attribution produced by our neural SDE is visibly sharper than the neural ODE attribution.

A natural question to ask is:

> *Can we theoretically and experimentally show that attributions corresponding to neural SDEs are more robust to small input perturbations than those obtained from neural ODEs?*

Brief Summary of Results. Our main result is theoretical and empirical demonstration that neural SDEs create more robust attributions than neural ODEs. We consider a neural network such that a change in the input $\epsilon_{\mathbf{x}}$ causes a change of $\epsilon_{\mathcal{F}}$ in the output layer of the neural SDE or ODE. In most applications, linear layers are then used for projecting to a different dimension such as the number of classes of the overall deep learning classification model. In this paper, we focus on the residual layers to study the incremental benefit of using neural SDEs, and the output of the SDE or ODE model is of the same dimension as that of the input.

If the evolution of the dynamical system is contracting in the neighborhood of an input $\mathbf{x}$, and $\epsilon_{\mathcal{F}}$ is less than $\epsilon_{\mathbf{x}}$, then the dynamics is inherently robust to perturbations in the input. But in most applications such as classification, the dynamics is expected to increase $\epsilon_{\mathcal{F}}$ with the increase in input perturbation helping create easy-to-learn classification boundaries between the inputs of different classes. This points to a natural trade-off between robustness and accuracy of deep learning models. While an increase in the ratio $\epsilon_{\mathcal{F}}/\epsilon_{\mathbf{x}}$ helps improve the model accuracy, it also indicates a reduced robustness as very small changes $\epsilon_{\mathbf{x}}$ in the input will produce large changes $\epsilon_{\mathcal{F}}$ in the output representation. The difference in the sum of attribution over the features of an input $\mathbf{x}$ and that over a perturbed input $\hat{\mathbf{x}}$ can be computed as the difference in the output representations when the attribution technique satisfies the completeness axiom [Sundararajan *et al.*, 2017]. We observe that robustness of the attribution for an input $\mathbf{x}$ appears to be at odds with model accuracy. This is not surprising since models with high accuracy are naturally expected to have multiple refined cues helping in making decisions, and a small change in the input can alter the cues responsible for the decision even if the decision itself remains unchanged. This presence of different set of fine-grained cues in high-accuracy models would generate different attributions over the input features.

3 Approach and Theoretical Results

We show that the attributions of neural stochastic differential equations are inherently more robust than the attributions of their deterministic counterparts. In Sec. 3.1, we briefly recapitulate the axiomatic definition of integrated gradients and then review the connection between residual networks with noise and neural stochastic differential equations (SDEs) in Section 3.2 before presenting our theoretical results.

3.1 Axiomatic Attributions

In this section, we discuss the axiomatic definition of attributions that will be used in subsequent sections to establish the robustness for neural SDEs. Attribution methods including integrated gradients and Shapley values often employ the notion of a baseline input $\mathbf{x}^b$; for example, the all dark image can be a baseline for images. The baseline can also be a set of random inputs where attribution is computed as an expected value.

Let the attribution for the j-th feature and output label i be denoted by $\mathcal{A}_j^i(\mathbf{x})$. The attribution for the j-th input feature depends on the complete input $\mathbf{x}$ and not just x_j. The treatment for each logit is similar, and so, we can drop the logit/class and denote the network output simply as $\mathcal{F}(\cdot)$ and attribution as $\mathcal{A}_j(\mathbf{x})$. One can readily make the following two assumptions on the DNN model and the attributions, which reflect that the model is well-trained and the attribution method is well-founded:

1. The attribution is dominated by the linear term. This is also an assumption made by attribution methods based on Shapley values such as Integrated Gradient [Sundararajan *et al.*, 2017] which define attribution as the path integral of the gradients of the DNN output with respect to that feature along the path from the baseline $\mathbf{x}^b$ to the input $\mathbf{x}$, that is,

$$\mathcal{A}_j^i(\mathbf{x}) = (\mathbf{x}_j - \mathbf{x}_j^b) \times \int_{\alpha=0}^{1} \partial_j \mathcal{F}^i(\mathbf{x}^b + \alpha(\mathbf{x} - \mathbf{x}^b)) d\alpha \tag{1}$$

where the gradient of i-th logit output of the model along the j-th feature is denoted by $\partial_j \mathcal{F}^i(\cdot)$.

2. Attributions are complete, that is, the following property is true for any input $\mathbf{x}$ and the baseline input $\mathbf{x}^b$:

$$\mathcal{F}(\mathbf{x}) - \mathcal{F}(\mathbf{x}^b) = \sum_{k=1}^{n} \mathcal{A}_k(\mathbf{x}) \text{ where } \mathbf{x} \text{ has } n \text{ features.} \tag{2}$$

Shapley value methods such as Integrated Gradient and DeepShap [Sundararajan *et al.*, 2017; Lundberg and Lee, 2017] satisfy this completeness axiom.

Our proof for robustness of attributions uses the completeness axiom satisfied by the attribution methods. In particular, we use this axiom to establish that the pathwise change in the sum of attributions $\sum_{k=1}^{n} (\mathcal{A}_k(\hat{\mathbf{x}}) - \mathcal{A}_k(\mathbf{x}))$ for two nearby inputs $\mathbf{x}$ and $\hat{\mathbf{x}}$ can be related to the pathwise change in the output representation $\mathcal{F}(\hat{\mathbf{x}}) - \mathcal{F}(\mathbf{x})$ of the neural ODE/SDE model.

3.2 Neural Stochastic Differential Equations

We connect with the recent literature on dynamical systems for neural networks to analyze the robustness of attributions of neural networks. Dynamical systems such as ordinary and stochastic differential equations can be used to model residual networks (ResNets). The evolution of these dynamical systems represents the process of inference in ResNets [Chen *et al.*, 2018] where the layers can be viewed as discretization of the continuous inference dynamics. This continuous dynamics view of inference and learning has been shown to improve efficiency as well as analyzability of the deep learning models [Chang *et al.*, 2017; Chen *et al.*, 2018; Lu *et al.*, 2018]. In this paper, we exploit this connection between dynamical systems and neural networks to theoretically study the robustness of neural network attributions using results from dynamical systems. We show that neural SDEs where noise is injected at different layers of the neural network produce more robust attributions than neural ODEs in Sec. 3.3.

A building block of a residual neural network [He *et al.*, 2016] with the residual mapping $R(X(i), W(i))$ can be described using the following equation:

$$X(i+1) = X(i) + R(X(i), W(i)) \tag{3}$$

Here, $X(i)$ is the input to the i^{th} residual network building block and $X(i+1)$ is the corresponding output that serves as an input to the next building block. The weights of the neural network layers in this ResNet building block are denoted by $W(i)$. In particular, $X(0)$ denoted the input $\mathbf{x}$ in this notation and the final output representation $\mathcal{F}$ of the residual neural network with depth T is denoted by $X(T)$.

After taking suitable limits, the evolution of the residual network can be described by the continuous evolution of a ResNet ordinary differential equation (ODE):

$$\frac{dX(t)}{dt} = G(X(t), W(t)) \tag{4}$$

Here, $G(X(t), W(t)) = \lim_{\delta t \to 0} \frac{R(X(t), W(t))}{\delta t}$ and $X(0)$ is the input to the neural network.

The residual network is extended to a stochastic residual network by adding a noise term $N(i)$:

$$X(i+1) = X(i) + R(X(i), W(i)) + N(i) \tag{5}$$

The neural dynamical system corresponding to such stochastic residual networks is obtained by generalizing to a stochastic differential equation [Wang *et al.*, 2019; Liu *et al.*, 2018; Liu *et al.*, 2020; Wang *et al.*, 2019]. Here, the noise is modeled by weighing a Brownian motion term $B(t)$ with a suitable diffusion coefficient $\sigma(X(t), t)$:

$$dX(t) = G(X(t), W(t))\, dt + \sigma(X(t), t)\, dB(t) \tag{6}$$

We show that the attributions corresponding to such neural stochastic differential equations are more robust to small perturbations in the input than those corresponding to the neural ODEs. We analyze the relative robustness of attributions for ResNets or neural ODEs and the stochastic ResNets or neural SDEs in Section 3.3.

3.3 Robustness of Attributions

In a neural ODE or SDE, the input $\mathbf{x}$ can be denoted by $X(0)$ and the output representation $\mathcal{F}(\mathbf{x})$ by $X(t)$. Given a perturbed input $\hat{\mathbf{x}}$, the inferred output representation $\mathcal{F}(\hat{\mathbf{x}})$ of the network may be different from its output representation $\mathcal{F}(\mathbf{x})$ to the original input $\mathbf{x}$.

Theorem 1. *Consider a neural ODE with final output representation $\mathcal{F}$ where $\epsilon_{\mathbf{x}}$ denotes a change in the input and $\epsilon_{\mathcal{F}}$ denotes the corresponding change in the output $\mathcal{F}$. Similarly, for a neural SDE, we denote the final output representation by $\tilde{\mathcal{F}}$, the change in input by $\tilde{\epsilon}_{\mathbf{x}}$, and the change in the final output representation by $\tilde{\epsilon}_{\mathcal{F}}$. The attribution over the inputs for the neural ODE and the neural SDE model is denoted by $\mathcal{A}$ and $\tilde{\mathcal{A}}$, respectively. The total change in attributions for a neural SDE is smaller than the total change in attributions for a neural ODE for the worst-case input $\mathbf{x}$ perturbed to $\hat{\mathbf{x}}$, that is,*

$$\|\sum_{k=1}^{n}(\mathcal{A}_k(\hat{\mathbf{x}})-\mathcal{A}_k(\mathbf{x}))\| \geq \|\sum_{k=1}^{n}\left(\tilde{\mathcal{A}}_k(\hat{\mathbf{x}})-\tilde{\mathcal{A}}_k(\mathbf{x})\right)\|$$

Proof. Given a neural ODE, one can compute a bound on the influence of a small change in the initial input $\mathbf{x}$ on the final output representation $\mathcal{F}$. $X(t)$ denotes the trajectory starting from the initial state $\mathbf{x}$. Let $\hat{C}_F \geq 0$ be a constant bounding the change in the dynamics, that is, for all $t \in [0, T]$, the following holds:

$$\|\mathcal{F}(\hat{\mathbf{x}})-\mathcal{F}(\mathbf{x})\| \leq \|\hat{\mathbf{x}}-\mathbf{x}\|\, e^{\hat{C}_F T} = \epsilon_{\mathbf{x}} e^{\hat{C}_F T} \tag{7}$$

From the completeness of attributions for the inputs $\mathbf{x}$ and $\hat{\mathbf{x}}$, we know the following:

$$\mathcal{F}(\mathbf{x})-\mathcal{F}(\mathbf{x}^b) = \sum_{k=1}^{n}\mathcal{A}_k(\mathbf{x}) \text{ and } \mathcal{F}(\hat{\mathbf{x}})-\mathcal{F}(\mathbf{x}^b) = \sum_{k=1}^{n}\mathcal{A}_k(\hat{\mathbf{x}}) \tag{8}$$

Subtracting attributions for $\mathbf{x}$ from those of $\hat{\mathbf{x}}$, we get the following result:

$$\epsilon_{\mathcal{F}} = \mathcal{F}(\hat{\mathbf{x}})-\mathcal{F}(\mathbf{x}) = \sum_{k=1}^{n}(\mathcal{A}_k(\hat{\mathbf{x}})-\mathcal{A}_k(\mathbf{x})) \tag{9}$$

We can similarly bound the change in attributions for the neural SDE. In order to compare a neural ODE and SDE for their relative robustness of attribution, we require them them to be trained with the same data and similar accuracy, except for the insertion of noise in SDEs. Due to this insertion of noise, the dynamics for the training data for neural SDEs will satisfy the following equation

$$\|\tilde{\mathcal{F}}(\hat{\mathbf{x}})-\tilde{\mathcal{F}}(\mathbf{x})\| \leq \|\hat{\mathbf{x}}-\mathbf{x}\|\, e^{(\hat{C}_F+N)T} \tag{10}$$

where N depends on the injected noise. In order to compare a neural ODE and a neural SDE, they need to be trained on the same data and for all data points (including the extreme points where the above inequality is actually an equality), the model needs to have the same accuracy. Thus,

$$\frac{\epsilon_{\mathcal{F}}}{\epsilon_{\mathbf{x}}} \geq \frac{\tilde{\epsilon}_{\mathcal{F}}}{\tilde{\epsilon}_{\mathbf{x}}} \tag{11}$$

because $e^{NT} \geq 1$ since N is positive for any added noise in training the neural SDE. □

4 Experimental Results

Our attribution analysis and model training is performed on a system with four NVIDIA V100 32GB GPUs using the ResNet-50, WideResNet-101-2 and ResNeXt-101 models [He *et al.*, 2016] on the ImageNet benchmark. The neural SDE models are trained by injecting a normalized noise into each layer of the residual network. We use the Adam optimizer and the cross-entropy loss in our experimental studies.

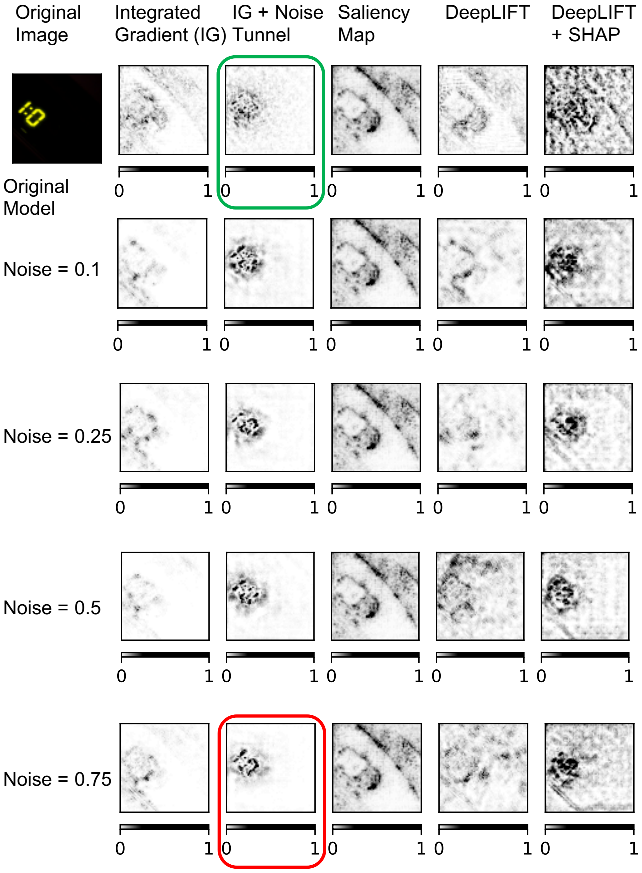

Figure 4: The top row illustrates the original input image of a digital clock and its attributions using integrated gradients, IG with a noise tunnel, saliency map, DeepLIFT and DeepLIFT SHAP. The visibly best attribution in the top row is obtained by integrated gradient with a noise tunnel and is indicated by a green box. Our neural SDE approach with various normalized magnitudes of noise applied to this attribution method creates a visibly better attribution in the explanation surrounded by a red box. Similar improvement is observed in DeepSHAP method that produces an initial reasonable attribution for the neural ODE. DeepSHAP produces a visibly sharper explanation for the neural SDE.

We use the Captum tool [Kokhlikyan *et al.*, 2020] for computing attributions and use a variety of different attribution methods including integrated gradients, Smoothgrad, DeepLIFT and DeepLIFT Shap to demonstrate the generality of our result. The improvement in attributions for neural SDEs is agnostic to the choice of attribution method. Further, we investigate qualitative improvement in the form of smooth robust attributions which are visibly sharper and well-defined, as well as improvement in more recently proposed quantitative metrics for attributions such as the sensitivity metric [Yeh *et al.*, 2019].

4.1 Qualitative Results

We analyzed attributions for 100 images from the ImageNet validation data set using both Neural ODEs and Neural SDEs for ResNet-50, WideResNet-101-2 and ResNeXt-101 models, and repeated the analysis 5 times to eliminate any significant statistical variations in our quantitative results. Each image has been analyzed using five different attribution methods: integrated gradients [Sundararajan *et al.*, 2017], integrated gradients with SmoothGrad [Smilkov *et al.*, 2017], Saliency Maps [Simonyan *et al.*, 2013], DeepLIFT [Shrikumar *et al.*, 2017] and DeepLIFT SHAP [Lundberg and Lee, 2017]. The Adam optimizer with a learning rate of 0.0001 and 100 epochs is used to train stochastic residual network models with a normalized injected noise of magnitudes $0.1, 0.25, 0.5$ and 0.75. Attributions are computed on inferences without any noise injection using the Captum library [Kokhlikyan *et al.*, 2020].

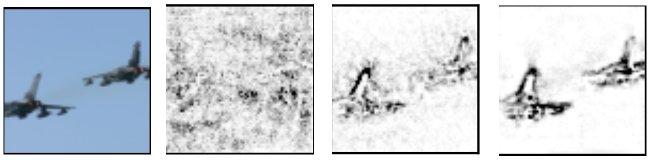

Figure 5: Neural SDE produces sharper integrated gradients than Neural ODEs. Left to Right: Original image, neural ODE + integrated gradients (IG), neural ODE + IG + noise tunnel, neural SDE + IG + noise tunnel (our approach).

We briefly highlight a few ResNet-50 examples in this section. Figure 5 shows the attributions obtained using integrated gradients and integrated gradients with a traditional noise tunnel [Smilkov *et al.*, 2017] for the input in the middle two images. The rightmost image is the attribution obtained using integrated gradients with a noise tunnel on the Neural SDE model with a normalized noise of 0.5; our neural SDE approach produces a visually sharper image.

The middle image of Figure 6 visualizes the attributions obtained by using DeepLIFT SHAP [Lundberg and Lee, 2017; Kokhlikyan *et al.*, 2020] on the neural ODE model. The rightmost figure is an outcome of applying the same attribution on a neural SDE model with a normalized noise of magnitude 0.5. The two planes are more clearly visible in the neural SDE explanation.

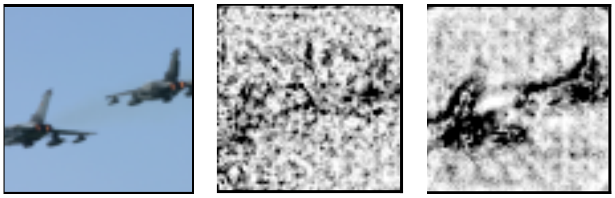

Figure 6: Neural SDE produces sharper DeepLIFT + SHAP attributions than those produced by Neural ODEs. Left to Right: Original image, neural ODE + DeepLIFT + SHAP, neural SDE + DeepLIFT + SHAP (our approach).

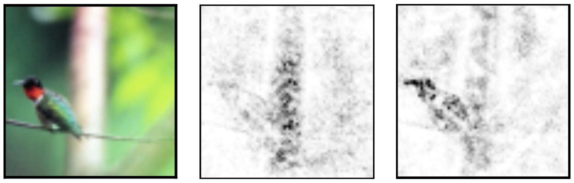

Figure 7: The DeepLIFT attributions of Neural SDE are visibly sharper than those produced by Neural ODEs. Left to Right: Original image, neural ODE + DeepLIFT, neural SDE + DeepLIFT (our approach).

Figure 7 illustrates the difference between the DeepLIFT attributions computed by neural SDEs and neural ODEs for the image of a humming bird from the ImageNet data set. The humming bird is clearly visible in the neural SDE attribution while the neural ODE attribution is diffused and spread over the tree and the bird in the image.

Figure 4 shows the impact of using Neural SDEs with different degrees of injected noise. The following two observations are relevant here:

1. A noise of small magnitude has an impact on the visual quality of attributions. However, we have not quantitatively studied the impact of the magnitude of the noise on the sensitivity metric [Yeh *et al.*, 2019].
2. The visual quality of the attributions improves as we increase the noise being injected. However, an increase in the magnitude of the noise also reduces the accuracy of the network. This trade-off in accuracy and robust attribution needs further investigation.

4.2 Quantitative Results

We quantitatively study the robustness of the computed attributions for both ResNet-50, WideResNet-101-2 and ResNeXt-101 models using the sensitivity metric [Yeh *et al.*, 2019] implemented in the Captum library [Kokhlikyan *et al.*, 2020]. We repeated our sensitivity analysis experiments 5 times on 100 correctly labeled but random images from the validation set of the ImageNet benchmark. We report the average values for ResNet-50 in Table 1 below.

		Sensitivity	
Reference	Attribution	ODE	SDE
[Simonyan *et al.*, 2013]	Saliency	0.5952	0.5510
[Sundararajan *et al.*, 2017]	IG	0.5788	0.4498
[Shrikumar *et al.*, 2017]	DeepLIFT	0.7498	0.6134
[Lundberg and Lee, 2017]	DeepSHAP	0.3566	0.3230

Table 1: Neural SDE produces a lower sensitivity metric than neural ODE for ImageNet benchmark using the ResNet-50 model.

Table 2 shows the average sensitivity metric obtained from five repetitions of our sensitivity analysis experiments on the WideResNet-101-2 model using both neural ODEs and neural SDEs. The sensitivity metrics [Yeh *et al.*, 2019] for neural SDEs are smaller than those for neural ODEs.

Reference	Attribution	Sensitivity	
		ODE	SDE
[Simonyan *et al.*, 2013]	Saliency	0.5972	0.5480
[Sundararajan *et al.*, 2017]	IG	0.5958	0.4940
[Shrikumar *et al.*, 2017]	DeepLIFT	0.8226	0.6674
[Lundberg and Lee, 2017]	DeepSHAP	0.3332	0.3236

Table 2: The sensitivity metrics for neural SDEs are smaller than those for neural ODEs. Our experiments used the WideResNet-101-2 model and the ImageNet benchmark.

Table 3 shows the sensitivity metrics obtained from the ResNeXt101 model using both neural ODEs and neural SDEs. Each value for sensitivity metric in the table is an average computed by performing the sensitivity analysis experiments 5 times on 100 ImageNet images.

Reference	Attribution	Sensitivity	
		ODE	SDE
[Simonyan *et al.*, 2013]	Saliency	0.6204	0.5570
[Sundararajan *et al.*, 2017]	IG	0.6068	0.4976
[Shrikumar *et al.*, 2017]	DeepLIFT	0.7972	0.7130
[Lundberg and Lee, 2017]	DeepSHAP	0.3692	0.3302

Table 3: Neural SDE produces a lower sensitivity metric than neural ODE for ImageNet using the ResNeXt101_32x8d model.

This, our experiments demonstrate that the neural SDEs have lower sensitivity metric than neural ODEs for each of the attribution methods: saliency maps, integrated gradients, Deep LIFT, and DeepLIFT SHAP, and this result generalizes across ResNet-50, WideResNet-101-2 and ResNeXt-101 models. As long as we have a residual neural network architecture that can be trained as a neural SDE, the attributions are robust, smoother, and qualitatively sharper.

5 Related Work

Model interpretability and attribution methods. A number of techniques [Lundberg and Lee, 2017; Sundararajan *et al.*, 2017; Li and Yu, 2015; Jha *et al.*, 2017; Jha *et al.*, 2019] for explaining deep neural networks have recently been proposed in the literature. These methods either provide a complete logical explanation for the output of the model or assign quantitative importance (attributions) to input features for a given model decision. Many of these methods are based on different analyses of the gradient of the predictor function with respect to the input [Simonyan *et al.*, 2013; Selvaraju *et al.*, 2017; Sundararajan *et al.*, 2017]. A number of attribution methods have been compared in [Adebayo *et al.*, 2018] and the sensitivity of these attributions to perturbations in the input has been studied in [Ghorbani *et al.*, 2019]. These studies highlight the need for additional methods that can make attributions of neural networks more robust to input perturbations. Our observation that neural SDEs produce more robust attributions with a smaller sensitivity score [Yeh *et al.*, 2019] is a step in this direction. The challenge of quantifying the improvement in robustness with the addition of noise during training of neural SDEs remains open.

Dynamical systems for neural networks. Dynamical systems models of neural networks have been the subject of several recent investigations with a particular emphasis on residual networks [Chang *et al.*, 2017; Weinan, 2017; Lu *et al.*, 2018; Tabuada and Gharesifard, 2020]. The theory of partial differential equations has been used to obtain dynamical system models of ResNets [Chang *et al.*, 2017; Weinan, 2017; Lu *et al.*, 2018]. Stochastic variants of residual neural networks have been described using neural stochastic differential equations [Wang *et al.*, 2019; Liu *et al.*, 2018; Liu *et al.*, 2020; Wang *et al.*, 2019]. Our main contribution in this paper is to leverage the connection between dynamical systems and neural networks to analyze the robustness and quality of attributions over the input features for a prediction by a deep learning model.

Connection between SmoothGrad and our work. SmoothGrad [Smilkov *et al.*, 2017] uses attributions over multiple noisy variants of an input image to generate visually sharp attributions. Our qualitative experimental results show that neural SDEs lead to SmoothGrad attributions that are visually sharper than SmoothGrad attributions obtained using neural ODEs. In our experiments, we use the noise tunnel implementation of SmoothGrad, as implemented in Captum [Kokhlikyan *et al.*, 2020]. Our approach can be seen as a generalization of the SmoothGrad [Smilkov *et al.*, 2017] approach where noise is injected not just in the input but in all the internal representations of the deep neural network.

6 Conclusions

We make three key observations in this paper. First, we mathematically analyze the attributions computed on neural SDEs trained using noise and show that these are more robust than attributions computed on the deterministic neural ODEs. Second, we experimentally show that this improvement in the computed attributions is agnostic to the choice of path-integral attribution method. We demonstrate improvement over state-of-the-art attribution methods including Saliency Maps, DeepLIFT SHAP and Integrated Gradients (see Fig. 1 through Fig. 4). Finally, we experimentally illustrate that attributions computed using neural SDEs have lower sensitivity scores than those computed using neural ODEs for ResNet-50, WideResNet-101-2 and ResNeXt-101 models.

Acknowledgements

The authors acknowledge support from the National Science Foundation awards #2113307, #1908471, and #1740079, the DARPA cooperative agreement #HR00112020002, ONR grant #N000142112332, and the U.S. Army Research Laboratory Cooperative Research Agreement W911NF-17-2-0196. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

[Adebayo *et al.*, 2018] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, pages 9525–9536, 2018.

[Chang *et al.*, 2017] Bo Chang, Lili Meng, Eldad Haber, Frederick Tung, and David Begert. Multi-level residual networks from dynamical systems view. *arXiv preprint arXiv:1710.10348*, 2017.

[Chen *et al.*, 2015] Yunjin Chen, Wei Yu, and Thomas Pock. On learning optimized reaction diffusion processes for effective image restoration. In *CVPR*, pages 5261–5269, 2015.

[Chen *et al.*, 2018] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *NeurIPS*, pages 6571–6583, 2018.

[Dombrowski *et al.*, 2019] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *NeurIPS*, pages 13589–13600, 2019.

[Ghorbani *et al.*, 2019] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *AAAI*, volume 33, pages 3681–3688, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Jha *et al.*, 2017] Susmit Jha, Vasumathi Raman, Alessandro Pinto, Tuhin Sahai, and Michael Francis. On learning sparse Boolean formulae for explaining AI decisions. In *NASA Formal Methods Symposium*, pages 99–114. Springer, 2017.

[Jha *et al.*, 2019] Susmit Jha, Tuhin Sahai, Vasumathi Raman, Alessandro Pinto, and Michael Francis. Explaining AI decisions using efficient methods for learning sparse Boolean formulae. *Journal of Automated Reasoning*, 63(4):1055–1075, 2019.

[Kim *et al.*, 2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, pages 2668–2677. PMLR, 2018.

[Kokhlikyan *et al.*, 2020] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

[Li and Yu, 2015] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015.

[Liu *et al.*, 2018] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *ECCV*, pages 369–385, 2018.

[Liu *et al.*, 2020] Xuanqing Liu, Tesi Xiao, Si Si, Qin Cao, Sanjiv Kumar, and Cho-Jui Hsieh. How does noise help robustness? explanation and exploration under the neural sde framework. In *CVPR*, June 2020.

[Lu *et al.*, 2018] Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *ICML*, pages 3276–3285. PMLR, 2018.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, pages 4765–4774, 2017.

[Selvaraju *et al.*, 2017] 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 *CVPR*, pages 618–626, 2017.

[Shrikumar *et al.*, 2017] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*, 2017.

[Simonyan *et al.*, 2013] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[Smilkov *et al.*, 2017] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[Sturmfels *et al.*, 2020] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328. JMLR. org, 2017.

[Tabuada and Gharesifard, 2020] Paulo Tabuada and Bahman Gharesifard. Universal approximation power of deep neural networks via nonlinear control theory. *arXiv preprint arXiv:2007.06007*, 2020.

[Wang *et al.*, 2019] Bao Wang, Zuoqiang Shi, and Stanley Osher. Resnets ensemble via the feynman-kac formalism to improve natural and robust accuracies. In *NeurIPS*, volume 32, pages 1657–1667, 2019.

[Weinan, 2017] E Weinan. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

[Xu *et al.*, 2020] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *CVPR*, pages 9680–9689, 2020.

[Yeh *et al.*, 2019] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in)fidelity and sensitivity of explanations. In *NeurIPS*, 2019.