Shaping Noise for Robust Attributions in Neural Stochastic Differential Equations

Sumit Kumar Jha[1], **Rickard Ewetz**[2], **Alvaro Velasquez**[3], **Arvind Ramanathan**[4], **Susmit Jha**[5]

[1] Computer Science Department, University of Texas at San Antonio, TX 78249
[2] Electrical and Computer Engineering Department, University of Central Florida, Orlando, FL 32816
[3] Information Directorate, Air Force Research Laboratory, Rome, NY 13441
[4] Data Science and Learning, Argonne National Laboratory, Lemont, IL, 60439
[5] Computer Science Laboratory, SRI International, Menlo Park, CA, 94709
sumit.jha@utsa.edu, rickard.ewetz@ucf.edu, alvaro.velasquez.1@us.af.mil, ramanathana@anl.gov, susmit.jha@sri.com

Abstract

Neural SDEs with Brownian motion as noise lead to smoother attributions than traditional ResNets. Various attribution methods such as saliency maps, integrated gradients, DeepSHAP and DeepLIFT have been shown to be more robust for neural SDEs than for ResNets using the recently proposed sensitivity metric. In this paper, we show that neural SDEs with adaptive attribution-driven noise lead to even more robust attributions and smaller sensitivity metrics than traditional neural SDEs with Brownian motion as noise. In particular, attribution-driven shaping of noise leads to 6.7%, 6.9% and 19.4% smaller sensitivity metric for integrated gradients computed on three discrete approximations of neural SDEs with standard Brownian motion noise: stochastic ResNet-50, WideResNet-101 and ResNeXt-101 models respectively. The neural SDE model with adaptive attribution-driven noise leads to 25.7% and 4.8% improvement in the SIC metric over traditional ResNets and Neural SDEs with Brownian motion as noise. To the best of our knowledge, we are the first to propose the use of attributions for shaping the noise injected in neural SDEs, and demonstrate that this process leads to more robust attributions than traditional neural SDEs with standard Brownian motion as noise.

Introduction

Computing the attribution of input features for a prediction made by a deep neural network (DNN) on an input has been extensively studied over the last few years (Simonyan, Vedaldi, and Zisserman 2013; Li and Yu 2015; Ribeiro, Singh, and Guestrin 2016; Kim et al. 2018; Jha et al. 2017; Sundararajan, Taly, and Yan 2017; Selvaraju et al. 2017; Shrikumar, Greenside, and Kundaje 2017; Lundberg and Lee 2017; Kapishnikov et al. 2019; Smilkov et al. 2017; Sturmfels, Lundberg, and Lee 2020). These attribution methods enable explaining decisions made by DNNs and consequently improve their interpretability. Attributions are also helpful in debugging DNNs for poor accuracy, generalizability, and fairness. These different attribution methods often rely on measuring the change in output with change in input features such as gradients with respect to the input or integrating gradients from a baseline to the actual input. While these approaches provide very compelling results, the computed attributions are still far from accurate. In this paper, we do not seek to develop yet another new approach to compute attributions but rather investigate whether we can train deep learning models that produce better attributions using existing attribution methods. We analyze the recently proposed Neural Stochastic Differential Equations (Neural SDEs) framework and present a new iterative training process that bootstraps the noise injected in each iteration using attributions computed over the previous iteration model. Such a Neural SDE model with adaptive attribution-driven noise produces qualitatively and quantitatively better attributions than the corresponding residual neural networks. Further, the attribution produced by adaptive attribution-driven shaping of noise in Neural SDEs are more robust than those produced by traditional Neural SDEs with Brownian motion as noise (Jha et al. 2021).

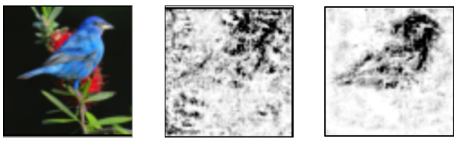

Figure 1: Integrated Gradients produces less noisy attributions for our adaptive attribution-driven Neural SDE model compared to the standard ResNet-50 model.

We formulate a new iterative approach in the technical approach section to train Neural SDEs that employs attributions from previous iterations to shape the injected noise in each iteration. This makes the model more robust over relevant features and also makes the attributions less noisy and more robust. We relate the theoretical results on the robustness of SDEs to the elimination of the saturation effects in attributions (Miglani et al. 2020; Smilkov et al. 2017) and alleviation of their sensitivity to irrelevant perturbations (Yeh et al. 2019; Ghorbani, Abid, and Zou 2019). We also empirically demonstrate the improved quality of attributions computed over Neural SDEs. We evaluate our approach on the ImageNet dataset in Table 1 and show that the attributions computed over such Neural SDEs with attribution-driven noise are consistently more robust to input perturbations

than those computed over neural SDEs with Brownian noise (Jha et al. 2021) across different methods for computing attributions. Figure 1 illustrates this improvement over a popular attribution computation method, Integrated Gradient (IG), for an example image. More qualitative results are presented in the experiment section. While these qualitative improvements are encouraging, systematically comparing attributions to evaluate them quantitatively is challenging due to the difficulty of knowing the ground truth. We address this challenge in the quantitative evaluation section where we summarize a list of quantitative metrics that formalize the expectations from accurate attributions, such as robustness to irrelevant perturbations, faithfulness and self-consistency with the model, and consistency with human annotations of foreground. Our results demonstrate that the Neural SDEs trained using attribution-based noise outperform attributions over ResNets across all these metrics for different attribution computation techniques. Further, neural SDEs with our attribution-driven noise are more robust and have lower sensitivity metric than recently proposed neural SDEs with Brownian motion noise (Jha et al. 2021). In summary, we make the following key contributions in this paper:

- We propose a new iterative approach for training Neural SDEs where attributions from the previous iteration are used in shaping the noise injected during the next iteration. This guided addition of noise on relevant features makes the attributions of such models more robust, that is, less sensitive to irrelevant perturbations compared to existing neural SDEs (Jha et al. 2021). See Table 1.
- We show that theoretical results on the robustness of SDEs help alleviate the saturation effect faced by attribution methods. We also empirically analyze this improvement and show that IG has a qualitatively different behavior over Neural SDEs compared to ResNets. See Fig. 3 and Fig. 4.
- We experimentally demonstrate that Neural SDEs with attribution-driven noise produce less noisy attributions than traditional Neural SDEs with standard Brownian motion noise using a number of different attribution methods: (i) Integrated Gradients, (ii) Noise Tunnel, (iii) Saliency Maps, (iv) DeepLIFT, and (v) DeepSHAP. See the section on experiments.
- We use 3 different quantitative evaluation metrics and demonstrate that the attributions over Neural SDEs with attribution-driven noise in comparison to traditional Neural SDEs with Brownian motion as noise show up to 19.4% relative improvement in robustness to input perturbations. Attribution-driven neural SDEs show better self-consistency with model output yielding a 4.8% relative increase in the SIC metric for the traditional Neural SDE model with Brownian noise and 25.7% relative increase with respect to ResNets. See Table 2.

Our results are a first step in studying the impact of shaping the injected noise in stochastic DNN architectures using the computed attributions on the qualitative and quantitative robustness of model attributions, which can aid in selecting the appropriate deep learning model to be used in applications where the explainability of decisions and the computation of accurate attributions is critical.

Related Work

Model interpretability and attribution methods. A number of explanation techniques (Simonyan, Vedaldi, and Zisserman 2013; Li and Yu 2015; Ribeiro, Singh, and Guestrin 2016; Kim et al. 2018; Sundararajan, Taly, and Yan 2017; Selvaraju et al. 2017; Shrikumar, Greenside, and Kundaje 2017; Lundberg and Lee 2017; Kapishnikov et al. 2019; Sturmfels, Lundberg, and Lee 2020) have recently been proposed in the literature that compute the relevant features or assign quantitative importance (attributions) to input features for a given decision by a DNN. Many of these methods are based on the gradient of the output with respect to an input feature (Simonyan, Vedaldi, and Zisserman 2013; Selvaraju et al. 2017; Sundararajan, Taly, and Yan 2017; Kapishnikov et al. 2019), and either directly use gradients, consider the product of gradients and activation, or integrate gradients over one or multiple paths. Different attribution methods are compared in (Adebayo et al. 2018). We do not seek to develop any new method to compute attributions in this paper. Instead, we demonstrate that a number of existing methods produce better attributions on Neural SDEs that are iteratively trained using attribution-based noise when compared with attributions on ResNets. The attribution methods are known to exhibit high sensitivity to irrelevant perturbations in the input which do not change the model's output but substantially change the attributions (Ghorbani, Abid, and Zou 2019). We show that the attributions on Neural SDEs are less sensitive and more robust to such perturbations. We theoretically and empirically show that Neural SDEs also alleviate the recently reported saturation problem of integrated gradients (Miglani et al. 2020). A number of methods have also been proposed to modify attribution methods to decrease the noise in attributions (Kapishnikov et al. 2019). The use of multiple noisy variants of an input during inference (Smilkov et al. 2017) and one-shot injection (Jha et al. 2021) of a fixed noise into all the layers of a neural SDE have been proposed in the literature. In contrast, we iteratively inject noise at each layer of the Neural SDE based on the attributions from previous iterations during training, which enhances the robustness of high-attribution features in the learned model. Our analysis of Neural SDE models for producing better attributions is complementary to the different attribution computation techniques.

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. The theory of partial differential equations has been used to obtain dynamical system models of ResNets (Chen, Yu, and Pock 2015; Chang et al. 2017; Sonoda and Murata 2017; Weinan 2017; Lu et al. 2018). Our work builds on the stochastic variants of residual neural networks modeled as Neural SDEs (Tzen and Raginsky 2019; Kidger et al. 2021; Wang, Shi, and Osher 2019; Liu et al. 2018; He, Rakin, and Fan 2019). To the best of our knowledge, we are the first to employ attributions to shape the noise used in training a neural SDE and show theoretically and experimentally that this leads to qualitatively and quantitatively better attributions. Our results will provide additional impetus to the study of Neural SDEs.

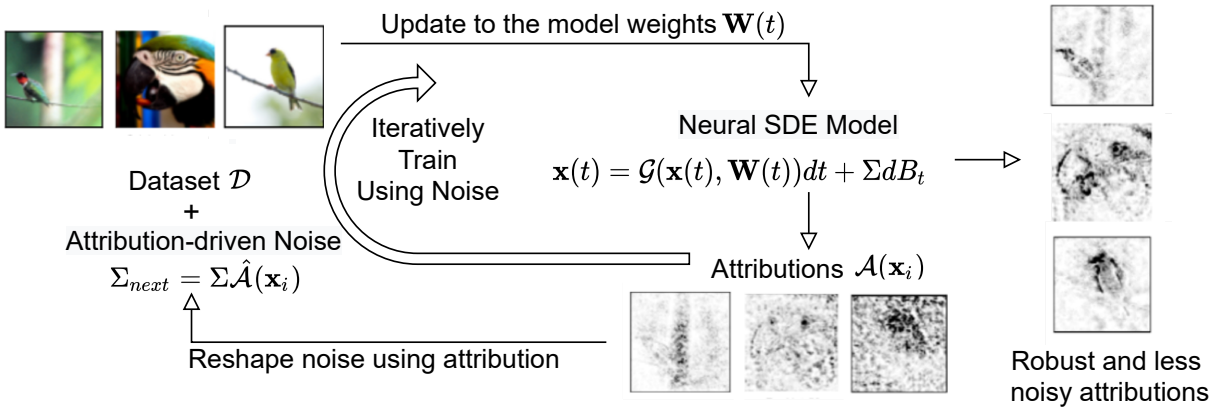

Figure 2: We use an iterative approach to train Neural SDEs using noise based on the attributions from the previous iteration. We start with uniform noise but iteratively reshape the noise to focus on features with high attribution. This makes the Neural SDE models more robust with respect to relevant features and also makes the attributions robust.

Technical Approach

The dynamics of ResNets has been described using partial differential equations (Chen, Yu, and Pock 2015; Chang et al. 2017; Sonoda and Murata 2017; Weinan 2017; Lu et al. 2018; Chen et al. 2018) where each building block of ResNets is modeled as one time-step of the dynamics. Formally, a residual building block $\mathbf{x}(i+1) = \mathcal{F}(\mathbf{x}(i), \mathbf{W}(i)) + \mathbf{x}(i)$ can be interpreted as the Euler discretization of a corresponding ordinary differential equation $\frac{d\mathbf{x}(t)}{dt} = \mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))$ such that $\mathcal{G}(\mathbf{x}(t), \mathbf{W}(t)) = \frac{\mathcal{F}(\mathbf{x}(t), \mathbf{W}(t))}{\delta t}$, where the input is $\mathbf{x}(0)$. Let $u(\mathbf{x}, t)$ be a function that is constant along the trajectory of this ODE. Then $u(\mathbf{x}, t)$ satisfies the following transport equation:

$$\frac{du(\mathbf{x},t)}{dt} = \frac{\partial u(\mathbf{x},t)}{\partial t} + \mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))\nabla u(\mathbf{x}, t) = 0 \quad (1)$$

The term $u(\mathbf{x}, t)$ serves as the classifier with the model output $u(\mathbf{x}, 0) = u(\mathbf{x}, 1)$, and the velocity field $\mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))$ encodes the architecture and weights of the DNN model. When $\mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))$ is very complex, $u(\mathbf{x}, t)$ can be very irregular and a small change in the input can change the classification, thereby making the DNN model not robust.

Training Neural SDEs With Attribution-based Noise Neural SDEs (Tzen and Raginsky 2019; Kidger et al. 2021; Wang, Shi, and Osher 2019; Liu et al. 2018; He, Rakin, and Fan 2019) have been recently proposed, and they provide a systematic approach for improving the robustness of ResNets. The key idea is to introduce an additional diffusion term $\frac{1}{2}\sigma^2\Delta u$, where Δ is the Laplace operator $\sum_i \frac{\partial^2}{\partial x_i^2}$. The corresponding convection-diffusion equation after introducing the diffusion term that captures robustness to small input perturbations is:

$$\frac{\partial u(\mathbf{x},t)}{\partial t} + \mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))\nabla u(\mathbf{x}, t) + \frac{1}{2}\sigma^2\Delta u = 0 \quad (2)$$

The above convection-diffusion equation can be solved using the Feynman-Kac formula over the following Itô process: $d\mathbf{x}(t) = \mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))dt + \Sigma dB_t$. The diffusion term makes the level sets of the transport equation more regular and hence, the classifier $u(\mathbf{x}, t)$ more robust as summarized in the following two theorems from the literature on stochastic dynamics that are also applicable to Neural SDEs (Ladyzhenskaia, Solonnikov, and Ural'tseva 1988; Wang, Shi, and Osher 2019).

Theorem 1. *If $\mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))$ is Lipschitz function in both $\mathbf{x}$ and t, the target classifier being learned is a compactly supported bounded function, and $0 < \sigma \leq 1$, then the solution $u(\mathbf{x}, t)$ for the convection-diffusion equation in Eqn. 2 satisfies $\|u(\mathbf{x}+\delta, 0) - u(\mathbf{x}, 0)\| \leq \alpha(\frac{\|\delta\|_2}{\sigma})^\beta$ for any small perturbation δ, where $\beta > 0$ and α depends on the infinity norm of $\mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))$.*

Thus, increasing the noise σ added during the training of the Neural SDE forces the learned classifier u to be more robust. A similar robustness theorem holds for the gradient $\nabla u(\mathbf{x}, t)$ of the classifier with respect to the input as summarized below.

Theorem 2. *If $\mathcal{G}(\mathbf{x}(t), \mathbf{W}(t)) \in C^1$ in both $\mathbf{x}$ and t, the target classifier being learned is a compactly supported bounded function, and $0 < \sigma \leq 1$, then the solution $u(\mathbf{x}, t)$ for the convection-diffusion equation in Eqn. 2 satisfies $\|\nabla u(\mathbf{x}, 1)\|_\infty \leq \alpha e^{-\sigma^2+\beta}$, where β depends on $\nabla\mathcal{G}$ and α depends on the infinity norm of the target classifier and its gradient.*

We can thus bound the gradient magnitude of the classifier model with respect to an input feature by adding more noise while training the Neural SDE. These theoretical results also motivated our use of attribution-based noise instead of uniform noise. Intuitively, the addition of noise smooths the learned classifier and regularizes its curvature. Our goal is to learn models that have better attribution without a significant loss in accuracy.

In traditional neural SDE models, one employs a scaled standard Brownian noise B_t while training the Neural SDE

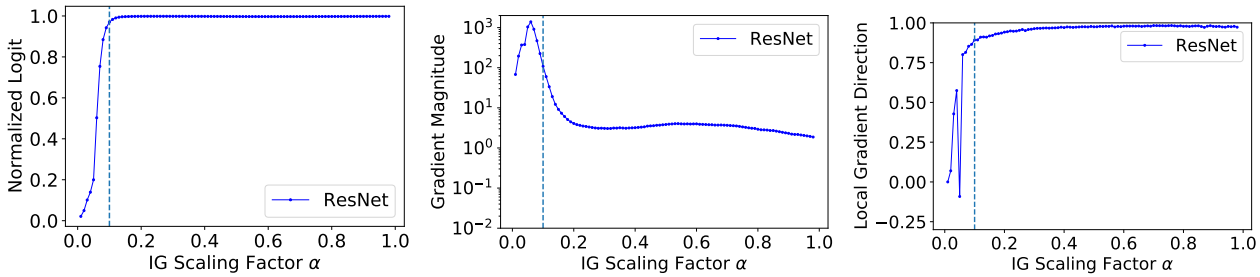

Figure 3: IG over ResNets: The left plot shows that the model reaches 95% of its final logit value around $\alpha = 0.1$. After this saturation of logit, the expectation is that the gradient magnitudes should either become small so that they do not contribute to the final IG, or the gradients are randomly aligned and hence, cancel out in the integration. But the center plot and the right plot show that the gradients remain significantly high after saturation, and their directions remain aligned in ResNets. Thus, the gradients after saturation significantly contribute to the Integrated Gradient in ResNets.

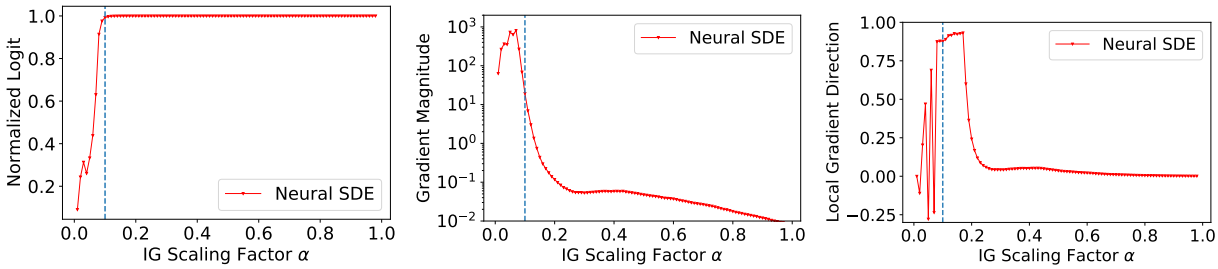

Figure 4: IG over Neural SDEs: The left plot shows that this model also reaches 95% of its final logit value around $\alpha = 0.1$. After this saturation of logit, the magnitude of the gradients in attribution-driven Neural SDE drops below 0.1 and does not significantly contribute to the final integration. The center figure shows the low magnitude of the gradients in Neural SDEs after reaching logit saturation, and the right figure shows that the average gradient alignment drops to zero, that is, the small gradients after saturation are randomly directed and will likely cancel each other out. Thus, the gradients after logit saturation do not significantly contribute to the Integrated Gradient in Neural SDEs which is the desirable behavior and reduces noise in attribution.

model: $d\mathbf{x}(t) = \mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))dt + \Sigma dB_t$. Here, Σ is not dependent on the attributions of the model.

In attribution-driven neural SDE approach $d\mathbf{x}(t) = \mathcal{G}(\mathbf{x}(t), \mathbf{W}(t))dt + \Sigma_{next} dB_t$, we use attribution to identify important features and inject noise $\Sigma_{next} = \Sigma\hat{\mathcal{A}}(\mathbf{x})$ based on the unit-normalized attribution $\hat{\mathcal{A}}(\mathbf{x})$. Given the attribution $\mathcal{A}(\mathbf{x})$, we obtain the unit-normalized attribution $\hat{\mathcal{A}}(\mathbf{x})$ by dropping the outliers in the computed attributions $\mathcal{A}(\mathbf{x})$ and normalizing the attribution values between -1 and 1.

Figure 2 illustrates this iterative approach which is repeated until the training loss converges. We initialize the noise to a Gaussian noise but over iterations, the noise is reshaped to make the high attribution features more robust. Traditional ResNets can be viewed as discretization of the inference in the Neural SDEs with zero noise, and this enables us to use the corresponding ResNets as a baseline for evaluation.

Saturation Effect of Integrated Gradient Recent results (Miglani et al. 2020; Smilkov et al. 2017) have shown that attribution techniques such as Integrated Gradient (IG) produce noisy observations due to bias in selection of baselines or gradient accumulation in saturated regions. The IG attribution of feature $\mathbf{x}_i$ using baseline $\mathbf{x}^b$ is given by $(\mathbf{x}_i - \mathbf{x}_i^b) \int_0^1 \frac{\partial \mathcal{F}(x_i^b + \alpha \cdot (\mathbf{x}_i - \mathbf{x}_i^b))}{\partial \mathbf{x}_i d\alpha}$. IG satisfies a set of axiomatic properties, including completeness which guarantees that the attributions sum to the difference between the model outputs on input and baseline.

A particularly perplexing observation of IG attributions over ResNets is that the contribution of the gradients in saturated regions of α where the model output changes minimally often dominate in magnitude and direction the gradients in the unsaturated regions where the model output changes substantially. This disproportionately larger impact of gradients in the saturated region produces noisy attribution. Splitting the integral and using attributions from the unsaturated region (Miglani et al. 2020) partly resolves this problem. We first demonstrate this problem with Integrated Gradient over ResNet models using the ImageNet dataset in Figure 3. We then illustrate how Neural SDE models with attribution-driven noise do not exhibit this limitation in Figure 4 and relate this to the theoretical results on the robustness of gradients. This alleviation of the saturation problem contributes to the overall improvement in the quality of attributions produced by attribution-driven Neural SDE models.

Experimental Results

We emphasize that our approach is not a new method for computing attributions; instead, it uses attribution-based iterative training of Neural SDEs to create models which produce more robust attributions than traditional Neural SDEs with Brownian motion noise (Jha et al. 2021) and also create better attributions than the corresponding ResNets (He et al. 2016). This improvement in interpretability of neural SDEs with attribution-driven noise is observed across a number of the state-of-the-art attribution methods. For evaluating our Neural SDE models with attribution-driven noise, we use both the corresponding Neural SDE models with Brownian noise and ResNets which are created by setting the noise to 0 and hence, form a fair baseline for comparison. In this section, we experimentally demonstrate that our attribution-driven Neural SDEs have quantitatively robust attributions than Neural SDEs with standard Brownian noise and the corresponding ResNets.

Our stochastic training and attribution analysis were performed on the ImageNet benchmark using 8 A100 GPUs with 40GB RAM. ImageNet training was performed using a learning rate of 0.0001 with a ReduceLROnPlateau scheduler, the noise constant $\sigma = 0.5$, and the Adam optimizer. Attribution analysis was performed on 1,000 examples from the ImageNet benchmark using the ResNet-50 model implemented in Pytorch. Training ResNet-50 on ImageNet using attribution-driven noise is 36.1% slower than training with Gaussian noise and is 38.7% slower than training without noise. The computation of attributions from a previous iteration can be computed in parallel with the training of the model to accelerate the training. Evaluating the efficacy of attribution is challenging due to the lack of direct ground truth for attribution. While human annotation of the qualitatively relevant and irrelevant part of an input is feasible, it is impractical to have ground truth of quantitative significance of different parts of the input. Thus, we consider a variety of different quantitative evaluation metrics in addition to qualitative comparisons.

Qualitative Evaluation Demonstrating Improvement Across Attribution Methods *Is the improvement in interpretability of neural SDE models trained using attribution-guided noise limited to specific attribution techniques, or does it generalize across attribution methods?* In this section, we use a few examples of qualitative results to illustrate that neural SDEs create more interpretable models, and different attribution methods such as DeepShap, IG, DeepLIFT and Noise tunnel, produce better attributions with neural SDEs than standard ResNets.

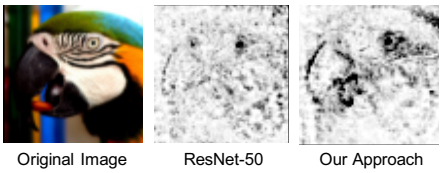

Figure 5: DeepSHAP Attributions on Macaw

DeepSHAP. Figure 5 shows an example of attribution obtained from our attribution-driven Neural SDE model using DeepSHAP – a combination of DeepLIFT and Shapley values implemented in the Captum library. DeepSHAP produces visually sharper and qualitatively better attributions using our Neural SDE model with attribution-driven noise than the standard ResNet-50 model. Our model has identified both the eye of the Macaw and its beak, while the ResNet-50 model has a more diffused attribution with some emphasis on the eye of the Macaw.

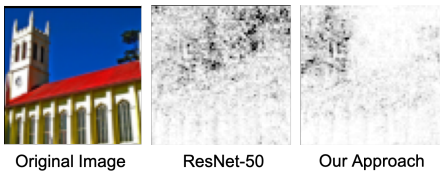

Figure 6: IG Attributions on Church

Integrated Gradient. Our Neural SDE model with attribution-driven noise produces visually sharper Integrated Gradient attributions than the standard ResNet-50 model. Fig. 6 shows how our stochastic model can focus on the watch tower with the spires (minarets) while the standard ResNet-50 model is focusing mostly on the blue sky in the center and right of the top of the image. This example also brings up the challenge of qualitative evaluation of attribution techniques and motivates our use of quantitative evaluation, which complements our qualitative analysis here.

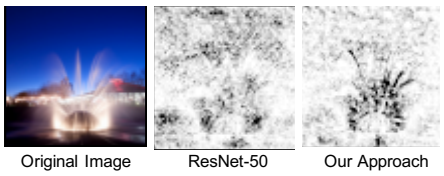

Figure 7: DeepLIFT Attributions on Fountain

DeepLIFT. Our Neural SDE model with attribution-driven noise leads to visually sharper DeepLIFT attributions than those produced by ResNet-50. Figure 7 shows an example of attribution obtained from ResNet-50 and our stochastic trained model using DeepLIFT as implemented in the Captum library. ResNet-50 has a very diffused attribution, while our stochastic model trained using attribution-driven noise is able to relatively focus on the water coming out of the fountain in the attributions, which is intuitively expected from an attribution approach.

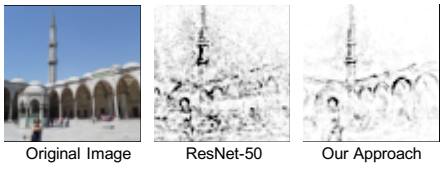

Figure 8: Noise Tunnel on Mosque

Model	Attribution	Sensitivity Metric		
		Standard	Noise	Attribution-driven Noise
ResNet-50	IG (Sundararajan, Taly, and Yan 2017)	0.576	0.450	0.420
	IG + NT (Smilkov et al. 2017)	1.036	0.983	0.866
	Saliency Map (Simonyan, Vedaldi, and Zisserman 2013)	0.596	0.551	0.478
	DeepLIFT (Shrikumar, Greenside, and Kundaje 2017)	0.729	0.613	0.554
	DeepSHAP (Lundberg and Lee 2017)	0.363	0.323	0.318
WideResNet-101	IG (Sundararajan, Taly, and Yan 2017)	0.561	0.494	0.461
	IG + NT (Smilkov et al. 2017)	1.433	1.426	1.408
	Saliency Map (Simonyan, Vedaldi, and Zisserman 2013)	0.577	0.548	0.501
	DeepLIFT (Shrikumar, Greenside, and Kundaje 2017)	0.777	0.667	0.643
	DeepSHAP (Lundberg and Lee 2017)	0.344	0.323	0.316
ResNeXt-101	IG (Sundararajan, Taly, and Yan 2017)	0.590	0.498	0.401
	IG + NT (Smilkov et al. 2017)	1.443	1.443	1.440
	Saliency Map (Simonyan, Vedaldi, and Zisserman 2013)	0.616	0.557	0.462
	DeepLIFT (Shrikumar, Greenside, and Kundaje 2017)	0.775	0.713	0.546
	DeepSHAP (Lundberg and Lee 2017)	0.379	0.330	0.321

Table 1: Lower sensitivity metric is desirable as it demonstrates robustness of attribution. For completeness, we consider different attribution methods and different ResNet architectures to ensure the quantitative improvement exhibited by the Neural SDE trained models generalize across different choices of models and attributions. The sensitivity of the corresponding ResNet models are in the third column, the sensitivity of the Neural SDE model trained without iterative attribution-driven noise is shown in the fourth column, and finally the sensitivity of our Neural SDE models is shown in the rightmost column.

Noise Tunnel (variant of SmoothGrad). Captum implements a variant of the SmoothGrad algorithm that is called the Noise Tunnel. A series of random perturbations of the inputs are used to refine the computed attributions. The performance of this approach on ResNet-50 on the mosque example in Figure 8 is good, but it includes lots of attributions up in the sky and on the ground. On the other hand, our Neural SDE model with attribution-driven noise focuses sharply on the minaret and the other architectural details of the mosque.

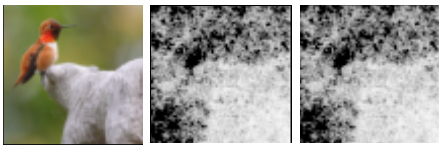

Figure 9: Saliency Map using just gradients does not produce visually sharp attributions on either our approach or ResNet-50.

Saliency Map. Saliency map is a foundational approach for explaining neural networks but does not compare well to more recent methods. As shown in Fig. 9, the saliency map does not produce visually sharp attributions using either our attribution-driven neural SDE approach or using ResNet-50. We include this as a failure case where our attribution-driven neural SDE approach also fails to improve the attributions of this particular approach; this could be due to the simple nature of the saliency map approach which directly uses the gradient of the input. This approach is known to exhibit a number of issues such as gradient saturation which could explain its failure on both approaches.

Quantitative Metrics for Evaluating Attributions over ResNets and Neural SDE models In this subsection, we focus on two questions.

- *What are the quantitative metrics to evaluate attributions beyond subjective qualitative judgements?*
- *Do our attribution-driven neural SDE models exhibit better interpretability and produce attributions that are quantitatively better than standard ResNet models?*

For quantitative evaluation of the attribution methods on standard ResNets, standard Neural SDEs, and those trained using our Neural SDE approach with attribution-driven noise, we use a number of metrics that formalize the following desiderata:

- **Robustness of attributions to input perturbations**: Perturbations to inputs that do not change the model output substantially should not change the attribution significantly. The computed attributions should be robust to such small perturbations of the input (Yeh et al. 2019; Ghorbani, Abid, and Zou 2019; Miglani et al. 2020).
- **Self-consistency with model output**: Attribution scores should be faithful to the model - removing the top or bottom features should lead to decrease or increase in the model's output (logit) for the class of the original input (Miglani et al. 2020; Sturmfels, Lundberg, and Lee 2020; Kapishnikov et al. 2019).
- **Consistency with weakly supervised localization**: The identified high-attribution features (pixels) of an input (image) should correspond to human-annotated foreground (Kapishnikov et al. 2019; Cong et al. 2018).

We next present results with each of these three quantitative metrics.

Model	Method	Reference	SIC
ResNet-50	Gradients	(Simonyan, Vedaldi, and Zisserman 2013)	0.510
ResNet-50	IG	(Sundararajan, Taly, and Yan 2017)	0.544
ResNet-50	IG + Noise Tunnel	(Smilkov et al. 2017)	0.590
NeuroSDE Standard Brownian Noise	IG	(Jha et al. 2021)	0.652
NeuroSDE Attribution-Driven Noise	IG	Our Approach	0.683

Table 2: The IG attributions over Neural SDE models trained using attribution-driven noise have higher SIC metric than IG attributions over traditional neural SDEs and standard ResNet-50 models.

Robustness of Attributions The robustness of attributions to perturbations that do not change the output model significantly can be computed over dataset $\mathcal{D}$ as the following sensitivity metric, where the magnitude of attribution-change is normalized by the original attribution magnitude:

$$\mathbb{S}_r(\mathcal{A}, \mathcal{D}) = \sum_{\mathbf{x} \in \mathcal{D}} \frac{\max_{\|\delta\|_\infty \leq r} \|\mathcal{A}(\mathbf{x}+\delta) - \mathcal{A}(\mathbf{x})\|_2}{\|\mathcal{A}(\mathbf{x})\|_2}$$

$$\text{such that } \forall \|\delta\|_\infty \leq r,\ F(x+\delta) = F(x)$$

This metric has been used previously in the literature (Yeh et al. 2019; Ghorbani, Abid, and Zou 2019; Miglani et al. 2020), and many state-of-the-art attributions (Sundararajan, Taly, and Yan 2017; Smilkov et al. 2017; Simonyan, Vedaldi, and Zisserman 2013; Shrikumar, Greenside, and Kundaje 2017; Lundberg and Lee 2017) have been found to be not robust for the standard ResNet models.

We demonstrate that models trained using our attribution-driven Neural SDE approach are more robust than both traditional Neural SDEs with Brownian noise and ResNets. Our approach is orthogonal to those improving robustness by averaging over multiple baselines, neighboring pixels, or different computation paths (Smilkov et al. 2017; Kapishnikov et al. 2019; Miglani et al. 2020), and can be combined with these approaches. For completeness, we consider different attribution methods and different architectures for the standard ResNets to ensure that the quantitative improvement exhibited by the Neural SDE trained models is not due to these choices. Our results are summarized in Table 1. Neural SDEs with attribution-driven noise have lower sensitivity across architectures and attribution methods.

Self-consistency with model output This self-consistency metric (Miglani et al. 2020; Sturmfels, Lundberg, and Lee 2020; Kapishnikov et al. 2019) is motivated by the need for the attribution method to be faithful to the model. For measuring this self-consistency, we adopt the Performance Information Curves (PIC) (Kapishnikov et al. 2019) where contents are re-introduced in a blurred (*bokeh*) version of the image to avoid sharp boundary effects (Dabkowski and Gal 2017) and the output is monitored. We use the proportion of the original input's label output or softmax score as the performance – such a PIC curve is also called the Softmax Information Curve (SIC) (Kapishnikov et al. 2019). The area under this SIC curve gives us a quantitative evaluation metric for the computed attributions. We evaluate the SIC metric on both the ResNet-50 model and our Neural SDE model trained using attribution-driven noise. We used 1,000 random images from the ImageNet validation data set. Our attribution-driven Neral SDE model has IG attributions with 4.75% higher SIC metric (Kapishnikov et al. 2019) than those obtained using traditional Neural SDEs and 25.7% higher SIC metric than those obtained from the standard ResNet-50.

Consistency with Weakly Supervised Localization Using human annotations such as those available from the PASCAL challenge, we identify a relevant part of the input denoted by G_F and an irrelevant background of the input denoted by G_B. We threshold the computed attributions to define binary masks over the input images that identify the high attribution relevant parts of the input R. The false positive rate and the true positive rate can be then computed as TPR $= \frac{|R \cap G_F|}{|G_F|}$, FPR $= \frac{|R \cap G_B|}{|G_B|}$, The area under the TPR and FPR curve gives the AUC score. We compute the AUC metric on 1,000 images from a subset of the ImageNet data set whose bounding boxes are available from the PASCAL VOC challenge. Our approach of injecting attribution-driven noise in Neural SDEs produces a slightly higher AUC score than the same IG approach applied to traditional Neural SDEs and the standard ResNet-50 model.

Conclusions

We developed a new iterative approach for training Neural SDEs where attributions from the previous iteration are used in shaping the noise in the next iteration. This shaping of noise using attributions while training Neural SDEs makes their attributions more robust than traditional SDEs with Brownian motion noise and ResNets. We show that SDEs alleviate the saturation effect faced by attribution methods and empirically demonstrate this. In future efforts, one can explore how such neural SDEs can lead to more robust confidence metrics (Jha et al. 2019) and enhance out-of-distribution detection algorithms (Kaur et al. 2022).

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, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. In *NeurIPS*, 9525–9536.

Chang, B.; Meng, L.; Haber, E.; Tung, F.; and Begert, D. 2017. Multi-level residual networks from dynamical systems view. *arXiv preprint arXiv:1710.10348*.

Chen, R. T.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. K. 2018. Neural ordinary differential equations. In *NeurIPS*, 6571–6583.

Chen, Y.; Yu, W.; and Pock, T. 2015. On learning optimized reaction diffusion processes for effective image restoration. In *CVPR*, 5261–5269.

Cong, R.; Lei, J.; Fu, H.; Cheng, M.-M.; Lin, W.; and Huang, Q. 2018. Review of visual saliency detection with comprehensive information. *IEEE Transactions on circuits and Systems for Video Technology*, 29(10): 2941–2959.

Dabkowski, P.; and Gal, Y. 2017. Real time image saliency for black box classifiers. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6970–6979.

Ghorbani, A.; Abid, A.; and Zou, J. 2019. Interpretation of neural networks is fragile. In *AAAI*, volume 33, 3681–3688.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, Z.; Rakin, A. S.; and Fan, D. 2019. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *CVPR*, 588–597.

Jha, S.; Ewetz, R.; Velasquez, A.; and Jha, S. 2021. On Smoother Attributions using Neural Stochastic Differential Equations. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 522–528. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Jha, S.; Raj, S.; Fernandes, S.; Jha, S. K.; Jha, S.; Jalaian, B.; Verma, G.; and Swami, A. 2019. Attribution-Based Confidence Metric For Deep Neural Networks. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jha, S.; Raman, V.; Pinto, A.; Sahai, T.; and Francis, M. 2017. On learning sparse Boolean formulae for explaining AI decisions. In *NASA Formal Methods Symposium*, 99–114. Springer.

Kapishnikov, A.; Bolukbasi, T.; Viégas, F.; and Terry, M. 2019. Xrai: Better attributions through regions. In *ICCV*.

Kaur, R.; Jha, S.; Roy, A.; Park, S.; Dobriban, E.; Sokolsky, O.; and Lee, I. 2022. iDECODe: In-distribution Equivariance for Conformal Out-of-distribution Detection.

Kidger, P.; Foster, J.; Li, X.; Oberhauser, H.; and Lyons, T. 2021. Neural SDEs as infinite-dimensional GANs. *arXiv preprint arXiv:2102.03657*.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, 2668–2677. PMLR.

Ladyzhenskaia, O. A.; Solonnikov, V. A.; and Ural'tseva, N. N. 1988. *Linear and quasi-linear equations of parabolic type*, volume 23. American Mathematical Soc.

Li, G.; and Yu, Y. 2015. Visual saliency based on multiscale deep features. In *CVPR*, 5455–5463.

Liu, X.; Cheng, M.; Zhang, H.; and Hsieh, C.-J. 2018. Towards robust neural networks via random self-ensemble. In *ECCV*, 369–385.

Lu, Y.; Zhong, A.; Li, Q.; and Dong, B. 2018. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *ICML*, 3276–3285. PMLR.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *NeurIPS*, 4765–4774.

Miglani, V.; Kokhlikyan, N.; Alsallakh, B.; Martin, M.; and Reblitz-Richardson, O. 2020. Investigating Saturation Effects in Integrated Gradients. *arXiv preprint arXiv:2010.12697*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *KDD*, 1135–1144.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *CVPR*, 618–626.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Sonoda, S.; and Murata, N. 2017. Double continuum limit of deep neural networks. In *ICML Workshop Principled Approaches to Deep Learning*, volume 1740.

Sturmfels, P.; Lundberg, S.; and Lee, S.-I. 2020. Visualizing the impact of feature attribution baselines. *Distill*, 5(1): e22.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *ICML*, 3319–3328. JMLR. org.

Tzen, B.; and Raginsky, M. 2019. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*.

Wang, B.; Shi, Z.; and Osher, S. 2019. ResNets Ensemble via the Feynman-Kac Formalism to Improve Natural and Robust Accuracies. In *NeurIPS*, volume 32, 1657–1667.

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

Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A. S.; Inouye, D. I.; and Ravikumar, P. 2019. On the (in) fidelity and sensitivity for explanations. *arXiv preprint arXiv:1901.09392*.