ExplainIt! A Tool for Computing Robust Attributions of DNNs

Sumit Jha[1] , **Alvaro Velasquez**[2] , **Rickard Ewetz**[3] , **Laura Pullum**[4] , **Susmit Jha**[5]
[1] Computer Science Department, University of Texas at San Antonio, San Antonio, TX 78249
[2] Information Directorate, Air Force Research Laboratory, Rome, NY13441
[3]ECE Department, University of Central Florida, Orlando, FL 32816
[4] Computer Science and Maths Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830
[5] Computer Science Laboratory, SRI International, Menlo Park, CA 94709

Abstract

Responsible integration of deep neural networks into the design of trustworthy systems requires the ability to explain decisions made by these models. Explainability and transparency are critical for system analysis, certification, and human-machine teaming. We have recently demonstrated that neural stochastic differential equations (SDEs) present an explanation-friendly DNN architecture. In this paper, we present ExplainIt, an online tool for explaining AI decisions that uses neural SDEs to create visually sharper and more robust attributions than traditional residual neural networks. Our tool shows that the injection of noise in every layer of a residual network often leads to less noisy and less fragile integrated gradient attributions. The discrete neural stochastic differential equation model is trained on the ImageNet data set with a million images, and the demonstration produces robust attributions on images in the ImageNet validation library and on a variety of images in the wild. Our online tool is hosted publicly for educational purposes.

1 Introduction

Artificial intelligence (AI), in particular deep neural networks (DNNs), have exceeded human-level performance in many applications such as computer vision and automatic control which form the integral components of many systems of social importance. Despite these successes, a key barrier to societal acceptance and wider adoption of DNNs is a lack of widely acceptable explanation methods that are robust, intuitive, and can explain their decisions to end-users. In this paper, we present `ExplainIt!`, an online tool for computing attributions of a neural network that can perform object recognition on images containing one of the 1,000 classes in the ImageNet dataset. The goal of the tool is to demonstrate that discrete approximations of neural SDEs lead to more robust and visually sharper explanations than traditional residual networks; this should encourage the exploration of neural SDEs for other domains that require robust explanations.

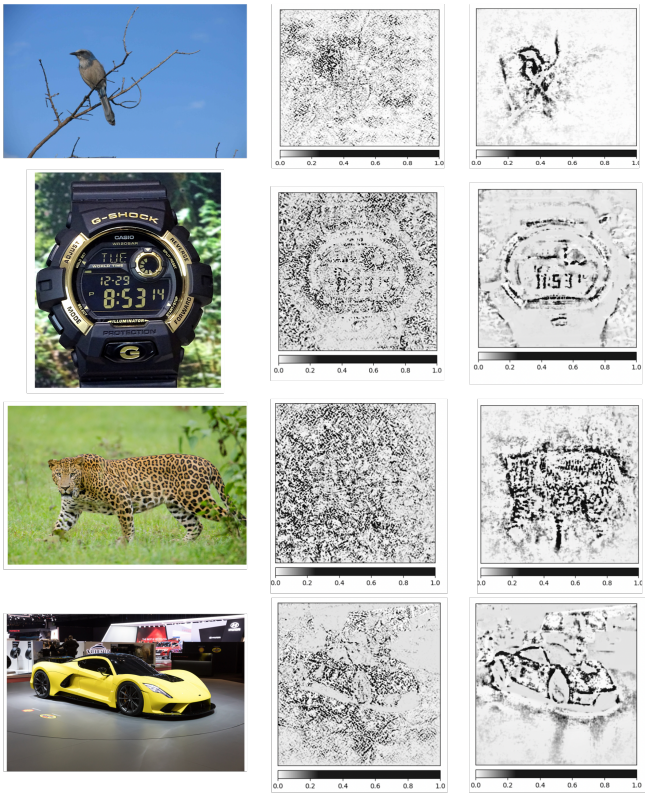

Figure 1: Integrated gradients explanations of images in the wild (left) using ResNet-50 (middle) and discrete approximation of Neural SDEs (right) from the `ExplainIt!` online tool.

2 Related Work

A lot of emphasis has been placed on developing new explanation methods for neural networks that either create a complete logical explanation or identify the features in an input that are crucial for the outcome of the neural network [Lundberg and Lee, 2017; Sundararajan *et al.*, 2017; Li and Yu, 2015; Yi *et al.*, 2016; Jha *et al.*, 2017; Jha *et al.*, 2019b]. The gradients of the neural network predictions with respect to the inputs and the model parameters, as well as mathematical constructs obtained from these gradients such as integrated gradients and Hessians, have been used to create approaches for building attribution algorithms [Sundararajan

et al., 2017]. These attributions have been shown to be susceptible to input perturbations [Ghorbani *et al.*, 2019], and this lack of robustness of attributions remains a challenge.

The literature on explaining AI decisions has a rich history [Simmons, 1988; Hammond, 1990; Swartout, 1983; Lane *et al.*, 2005; Core *et al.*, 2006]. Early methods on explaining AI decisions focused on planning problems, rule-based and expert systems. These foundational methods form the basis of many of the modern approaches being applied to deep neural networks. However, the need to explain decisions has now become more acute as neural networks have become huge in size and are being deployed in safety-critical settings.

3 Approach

Our goal is not to create yet another new explanation method for neural networks. Instead, we are investigating DNN architectures that produce more robust explanation across different attribution methods [Jha *et al.*, 2021; Jha *et al.*, 2022]. As shown in Fig. 2, our approach is based on injecting noise in every layer of the residual network as a discrete approximation to a neural stochastic differential equation (neural SDEs).

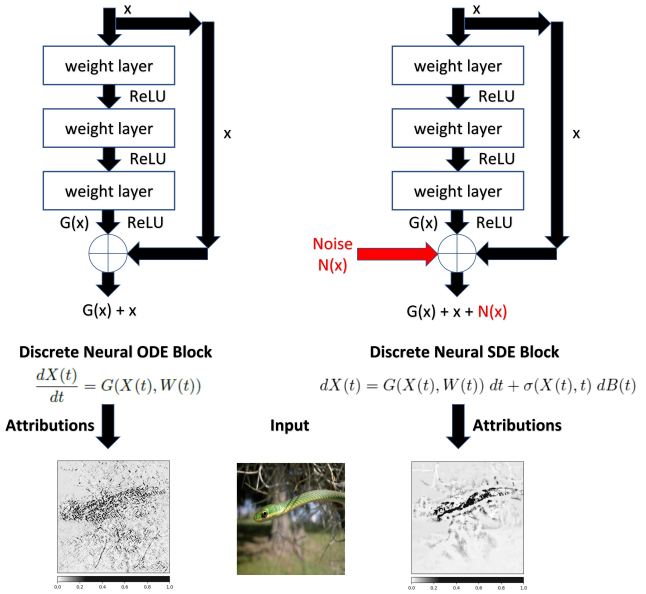

Figure 2: Our approach inserts noise in every layer of the neural network and this leads to more robust as well as visually sharper attributions. The approach is based on our earlier work [Jha *et al.*, 2021].

The following equation describes the elementary building block [He *et al.*, 2016] of a residual neural network (ResNet) with the residual mapping $R(X(i), W(i))$:

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

Here, the input internal representation to the i^{th} residual network building block is described by the notation $X(i)$. Similarly, $X(i+1)$ describes the corresponding output internal representation that then serves as an input feeding into the next building block of the residual network.

In particular, $X(0)$ denotes the input to the neural network in this notation and the final output of the residual neural network with depth T is denoted by $X(T)$ in this framework. The notation $W(i)$ is used to describe the weights associated with the i^{th} residual layer in the residual neural network.

The discrete model can then be generalized to a continuous model by taking suitable limits. The evolution of the residual neural network can be described by the following ordinary differential equation (ODE):

$$\frac{dX(t)}{dt} = G(X(t), W(t)) \tag{2}$$

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 can naturally be extended to a stochastic residual network by adding a noise term $N(i)$:

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

In such a stochastic network, a noise $N(i)$ is inserted into the internal representation of every layer of the neural network.

A stochastic differential equation [Wang *et al.*, 2019; Liu *et al.*, 2018; Liu *et al.*, 2020; Wang *et al.*, 2019] describes the dynamics of a continuous generalization of such stochastic residual neural networks:

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

Here, the noise inserted into the internal representation at every layer is mathematically described by multiplying a Brownian motion term $B(t)$ with a suitable diffusion coefficient $\sigma(X(t), t)$:

The addition of a small amount of noise into the internal representation of neural networks leads to smoother attributions for neural networks, as shown in our earlier work [Jha *et al.*, 2021]. Besides visual sharpness, we measure the robustness of the attributions by computing the change in attributions as the input is perturbed in a relatively small hypersphere of a fixed radius [Yeh *et al.*, 2019].

As shown in Table 1, neural SDEs produce robust explanations with lower sensitivity metrics [Yeh *et al.*, 2019] than those produced by the standard ResNet-50 model. Our work builds on prior work on the dynamical systems models of DNNs, such as neural ODEs and neural SDEs which have been investigated over the last few years [Chen *et al.*, 2015; Sonoda and Murata, 2017; Weinan, 2017; Lu *et al.*, 2018]. Stochastic variants of residual neural networks have also been described using neural SDEs [Wang *et al.*, 2019; Liu *et al.*, 2020; Wang *et al.*, 2019].

Attribution Approach	Sensitivity	
	ResNet-50	Neural SDE
Saliency	0.5952	0.5510
Integrated Gradients	0.5788	0.4498
DeepLIFT	0.7498	0.6134
DeepSHAP	0.3566	0.3230

Table 1: Neural SDE produces higher robustness than the ResNet-50 on the ImageNet benchmark from [Jha *et al.*, 2021].

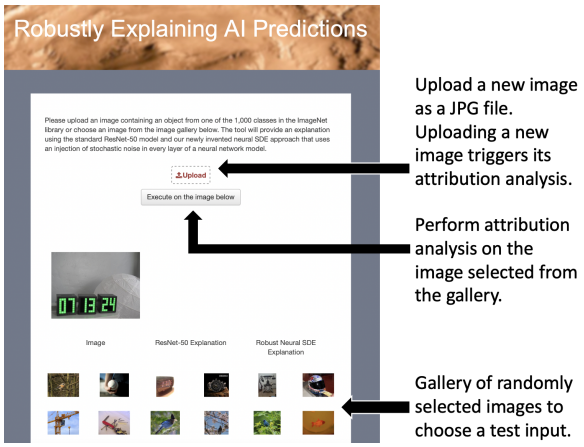

Figure 3: The interface of our tool `ExplainIt!` for attribution analysis of images using ResNet-50 and Neural SDE models.

4 Overview of the Online `ExplainIt!` Tool

We present a brief overview of the design and interface of the `ExplainIt!` tool hosted at https://explainit.sumitkumarjha.com

4.1 Demonstration

The publicly available interface of our tool (see Figure 3) has the option to select one of the images from the image gallery at the bottom of the page or upload a new image representing an object from one of the 1,000 ImageNet classes as a JPEG file. The uploaded input image can be a 3-channel image of any size. If an image is selected from the gallery, the user needs to press the "execute" button. If an image is uploaded by the user, it automatically triggers the attribution analysis of the image. The integrated gradient attributions of the image using a ResNet-50 model and the neural SDE approach are produced in less than a minute. Further technical details of our approach are presented in [Jha *et al.*, 2021].

4.2 Demonstration Interface

The screenshot of the demo is shown in Fig. 3. The user interface shown receives the input image from the user and displays the integrated gradient attributions from the ResNet-50 and discrete approximations of neural SDE models.

Data from the Wild

Our online tool performs well on images obtained in the wild. A small set of four images, their attributions computed by ResNet-50, and attributions of neural SDEs are shown in Figure 1. The tool allows users to upload additional images in the wild and observe the explanations.

ImageNet Validation Data Set

The image gallery in the online tool consists of images from the ImageNet data set. The images in the gallery were chosen to include scenarios where the predictions by the Neural SDE approach or the ResNet-50 approach are not accurate. In general, neural SDEs and their discrete approximations may produce lower test accuracy than ResNets but produce higher robustness in their attributions.

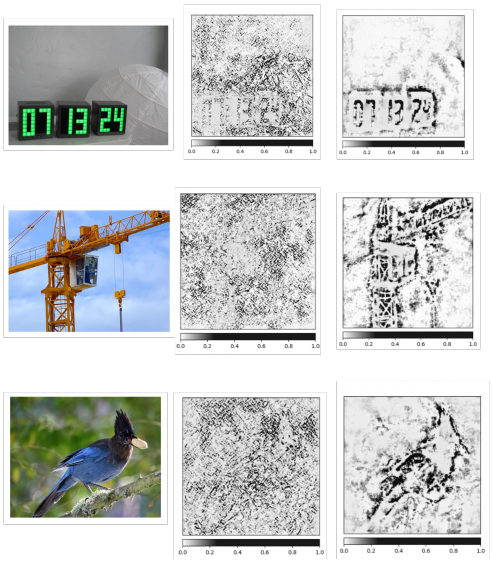

Figure 4: Integrated Gradient attributions for images from the ImageNet validation library. Images from the left to the right are inputs, ResNet-50 attributions, and discrete neural SDE attributions.

4.3 Design & Tool Architecture

The `ExplainIt!` tool is based on an HTML/CSS front-end hosted at https://explainit.sumitkumarjha.com that is built using a third-party Anvil open-source cloud engine. The front-end communicates the input image to a system with 8 80GB A100 GPUs, 256 3.3GHz cores and 2.2TB RAM that performs attribution analysis using both ResNet-50 and the Neural SDE models [Jha *et al.*, 2021]. While our approach can use any attribution method, the tool demo currently uses Integrated Gradient [Sundararajan *et al.*, 2017] as the attribution method. The generated explanations are returned to the front-end website for display to the end-user.

5 Conclusions

The `ExplainIt!` online tool allows a variety of users with different levels of AI expertise to perform attribution analysis of images using standard ResNet-50 and Neural SDE models without the need to access specialized hardware. Access to this online tool will raise awareness among AI users about the need to create models with more robust explanations. Better attributions can lead to the design of more precise confidence metrics [Jha *et al.*, 2019a], efficient automated synthesis [Jha and Raman, 2016], and explanation methods for continuous models [Cramer *et al.*, 2022]. Future versions of the `ExplainIt!` tool will include a choice of multiple attribution approaches.

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.

References

[Chen *et al.*, 2015] Yunjin Chen, Wei Yu, and Thomas Pock. On learning optimized reaction diffusion processes for effective image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5261–5269, 2015.

[Core *et al.*, 2006] Mark G Core, H Chad Lane, Michael Van Lent, Dave Gomboc, Steve Solomon, Milton Rosenberg, et al. Building explainable artificial intelligence systems. In *AAAI*, pages 1766–1773, 2006.

[Cramer *et al.*, 2022] Eike Cramer, Felix Rauh, Alexander Mitsos, Raúl Tempone, and Manuel Dahmen. Nonlinear isometric manifold learning for injective normalizing flows. *arXiv preprint arXiv:2203.03934*, 2022.

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

[Hammond, 1990] Kristian J Hammond. Explaining and repairing plans that fail. *Artificial intelligence*, 45(1-2):173–228, 1990.

[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 and Raman, 2016] Susmit Jha and Vasumathi Raman. Automated synthesis of safe autonomous vehicle control under perception uncertainty. In *NASA Formal Methods Symposium*, pages 117–132. Springer, 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.*, 2019a] Susmit Jha, Sunny Raj, Steven Fernandes, Sumit K Jha, Somesh Jha, Brian Jalaian, Gunjan Verma, and Ananthram Swami. Attribution-based confidence metric for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 11826–11837, 2019.

[Jha *et al.*, 2019b] 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.

[Jha *et al.*, 2021] Sumit Jha, Rickard Ewetz, Alvaro Velasquez, and Susmit Jha. On smoother attributions using neural stochastic differential equations. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 522–528, 2021.

[Jha *et al.*, 2022] Sumit Jha, Rickard Ewetz, Alvaro Velasquez, Arvind Ramanathan, and Susmit Jha. Shaping noise for robust attributions in neural stochastic differential equations. *36th AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[Lane *et al.*, 2005] H Chad Lane, Mark G Core, Michael Van Lent, Steve Solomon, and Dave Gomboc. Explainable artificial intelligence for training and tutoring. Technical report, University of Southern California, 2005.

[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 *Proceedings of the European Conference on Computer Vision (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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (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 *International Conference on Machine Learning*, pages 3276–3285. PMLR, 2018.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[Simmons, 1988] Reid G Simmons. A theory of debugging plans and interpretations. In *AAAI*, pages 94–99, 1988.

[Sonoda and Murata, 2017] Sho Sonoda and Noboru Murata. Double continuum limit of deep neural networks. In *ICML Workshop Principled Approaches to Deep Learning*, volume 1740, 2017.

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

[Swartout, 1983] William R Swartout. Xplain: A system for creating and explaining expert consulting programs. *Artificial intelligence*, 21(3):285–325, 1983.

[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 *Advances in Neural Information Processing Systems*, 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.

[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. *Advances in Neural Information Processing Systems*, 32:10967–10978, 2019.

[Yi *et al.*, 2016] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *ECCV*, pages 467–483. Springer, 2016.