Resilience of Bayesian Layer-Wise Explanations under Adversarial Attacks

Ginevra Carbone
Department of Mathematics and Geosciences, University of Trieste, Trieste, Italy
ginevra.carbone@phd.units.it

Luca Bortolussi
Department of Mathematics and Geosciences, University of Trieste, Trieste, Italy;
Modeling and Simulation Group, Saarland University, Saarland, Germany
luca.bortolussi@gmail.com

Guido Sanguinetti
School of Informatics, University of Edinburgh, Edinburgh, United Kingdom;
SISSA, Trieste, Italy
gsanguin@sissa.it

Abstract

We consider the problem of the stability of saliency-based explanations of Neural Network predictions under adversarial attacks in a classification task. Saliency interpretations of deterministic Neural Networks are remarkably brittle *even when the attacks fail*, i.e. for attacks that do not change the classification label. We empirically show that interpretations provided by Bayesian Neural Networks are considerably more stable under adversarial perturbations of the inputs and even under direct attacks to the explanations. By leveraging recent results, we also provide a theoretical explanation of this result in terms of the geometry of the data manifold. Additionally, we discuss the stability of the interpretations of high level representations of the inputs in the internal layers of a Network. Our results demonstrate that Bayesian methods, in addition to be more robust to adversarial attacks, have the potential to provide more stable and interpretable assessments of Neural Network predictions.

Introduction Deep Neural Networks (DNNs) are the core engine of the modern AI revolution. Their universal approximation capabilities, coupled with advances in hardware and training algorithms, have resulted in remarkably strong predictive performance on a variety of applications. However, the story of DNNs is tempered with a number of potentially very serious drawbacks which are somehow the natural flip side of dealing with extremely flexible and complex models. The first such drawback is the black box nature of DNNs: their expressivity and training on large data sets empirically results in very strong predictive power, but in general it does not provide any intuition about the possible explanations underlying the decisions. A second major drawback of DNN predictions is their vulnerability to adversarial attacks: empirically, it has been observed in many applications that well chosen infinitesimal changes in inputs can produce catastrophic changes in output [8], leading to paradoxical classifications and a clear problem in any application to safety critical systems.

In this paper, we argue theoretically and empirically that these two problems are interlinked, and that therefore solutions that ameliorate resilience against adversarial attacks will also lead to more stable and reliable interpretations. We work within the framework of saliency explanations, which attempt to interpret post-hoc DNN decisions by apportioning a relevance score to each input feature for each

data point. Specifically, we use the popular Layer-wise Relevance Propagation (LRP) [3], whose saliency interpretations are well known to be unstable under perturbations of the inputs [7, 9, 1, 16]. Our results confirm that the LRP robustness of deterministic DNN predictions is remarkably low even when the adversarial attack fail to change the overall classification of the data point, i.e. that LRP interpretations are *less robust* than actual classifications. Recently, [4] suggested that a Bayesian treatment might ameliorate these stability problems. Considerations on the geometry of LRP [2] suggest that the observed lack of robustness might be imputable to large gradients of the prediction function in directions orthogonal to the data manifold. We expand on such a point of view, integrating it with a theoretical analysis in a suitably defined large-data limit [5, 13, 6, 11], and leveraging recent results from [5] about the robustness of BNNs to gradient based adversarial attacks. Specifically, we prove that Bayesian training of the DNNs in the large-data and overparametrized limit induces a regularizing effect which naturally builds robust explanations. We empirically validate this claim in a variety of settings.

Methodology We consider the stability of saliency interpretations under targeted adversarial attacks that aim to change the classification under perturbations of the input. We introduce a novel notion of *k-LRP robustness* of relevance heatmaps to adversarial attacks and use this measure to assess how adversarial perturbations of the inputs affect the explanations.

Definition. *Let x be an image with relevance heatmap $R(x, w)$ and $\tilde{x}$ an adversarial perturbation with relevance heatmap $R(\tilde{x}, w)$. Let $\mathit{Top}_k(R)$ denote the pixel indexes corresponding to the top k percent most relevant pixels in the absolute value of a heatmap R. The* k-LRP robustness *of x w.r.t. the attack $\tilde{x}$ is*

$$k\text{-}\mathit{LRP}(x, \tilde{x}, w) := |\mathit{Top}_k(R(x, w)) \cap \mathit{Top}_k(R(\tilde{x}, w))|/k. \tag{1}$$

In other words, the $\text{Top}_k(R)$ pixels have a strong positive or negative impact on classification and k-LRP$(x, \tilde{x}, w)$ is the fraction of common most relevant pixels for x and $\tilde{x}$ in the top $k\%$. We analyse the behaviour of LRP representations in the internal layers of the network, thus we also extend the computation of LRP heatmaps to any feature representation of the input x at a learnable layer $l \in \mathbb{N}$ and denote it as k-LRP$(x, \tilde{x}, w, l)$. The notion of LRP robustness can be naturally generalised to the Bayesian setting using the concept of Bayesian model averaging. Hence, the LRP heatmap of a BNN is computed as the average of all the deterministic heatmaps from the ensemble: $\mathbb{E}_{p(w|D)}\big[k\text{-LRP}(x, \tilde{x}, w, l)\big]$.

To better conceptualise the impact of a Bayesian treatment on LRP robustness, it is convenient to consider the thermodynamic limit of infinite data and infinite expressivity of the network, as formalised in [6, 11, 13]. The main ingredients are the data manifold $\mathcal{M}_D$, a piecewise smooth submanifold of the input space where the data lie, and the true input/output function, which is assumed to be smooth and hence representable through an infinitely wide DNN. Because the data manifold is assumed to be piecewise smooth, it is possible to define a tangent space to the data manifold almost everywhere, and therefore to define two operators $\nabla_x^{\perp}$ and $\nabla_x^{\parallel}$ which define the gradient along the normal and tangent directions to the data manifold $\mathcal{M}_D$ at a point x of a function defined over the whole input space. In the thermodynamic limit, the DNN function $f(x, w)$ coincides with the true function everywhere on the data manifold, and therefore the tangent gradient of the loss function is identically zero. The normal gradient of the loss, however, is unconstrained by the data, and, particularly in a high dimensional setting, might achieve very high values along certain directions, creating therefore weaknesses that may be exploited by an adversarial attacker. The tangent components of the gradient of the prediction function will coincide with the gradients of the true function, and therefore represent directions of true sensitivity of the decision function which are correctly recognised as relevant. However, such directions might be confounded or dwarfed by normal gradient components, which create directions of apparent relevance which, by construction, are targeted by gradient-based adversarial attacks. In the following Theorem we prove that BNNs in the thermodynamic limit will only retain relevant directions along the data manifold, which correspond to genuine directions of high relevance.

Theorem. *Let $\mathcal{M}_D \subset \mathbb{R}^d$ be an a.e. smooth data manifold and let $f(x, w)$ be an infinitely wide Bayesian neural network, trained on $\mathcal{M}_D$ and at full convergence of the training algorithm. Let $p(w|D)$ be the posterior weight distribution and suppose that the prior distribution $p(w)$ is uninformative. In the limit of infinite training data, for any $x \in \mathcal{M}_D$, $\mathbb{E}_{p(w|D)}[\nabla_x^{\perp} f(x, w)] = 0$.*

Therefore, the orthogonal component of the gradient of the prediction function vanishes in expectation under the posterior weight distribution and Bayesian averaging of the relevance heatmaps naturally builds explanations in the tangent space $T_x\mathcal{M}_D$.

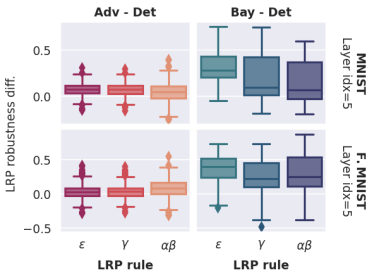

Figure 1: LRP robustness differences for FGSM attacks computed on 500 test points using Epsilon, Gamma and Alpha-Beta rules on the Top_{20} pixels, for adversarially trained networks (left) and BNNs (right). BNNs are trained with HMC and tested using 100 posterior samples.

Experimental Results Our first significant result is that Bayesian explanations are more robust under attacks than deterministic architectures. For multiple data sets, attacks, training techniques (SGD training, adversarial training [8, 10], Bayesian inference) and approximate inference methods (Hamiltonian Monte Carlo [12], Variational Inference [14]), LRP robustness scores are considerably higher than their deterministic counterparts. Adversarially trained networks have low LRP robustness compared to BNNs: this suggests empirically that the components of the gradient that are normal to the data manifold (and are therefore the ones likely to be changed in an attack) are often major contributors to the relevance in DNN. The experiments confirm that Bayesian explanations are more stable across multiple LRP rules, gradient-based adversarial attacks and saliency attacks, also in the internal layers (Fig. 1).

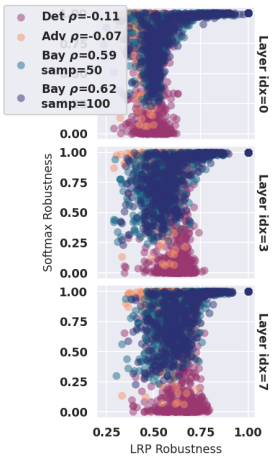

Figure 2: LRP vs softmax robustness on MNIST dataset against FGSM attack. ρ denotes the correlation coefficient. LRP Robustness is computed with the Epsilon rule on the 20% most relevant pixels.

A simple explanation for the improved LRP robustness of BNNs lies in the fact that BNNs are provably immune to gradient-based attacks [5]. Therefore, one might argue that the stability of the LRP is a trivial byproduct of the stability of the classifications. To explore this question more in depth, we consider the relationship between the LRP robustness of a test point (stability of the explanation) and its softmax robustness (resilience of the classification against an attack). Fig. 2 shows scatterplots of these two quantities for deterministic, adversarially trained and Bayesian NNs. Deterministic explanations are weak against adversarial perturbations even when their softmax robustness is close to 1. Therefore, even in the cases where the classification is unchanged, deterministic saliency heatmaps are fragile. Bayesian models, instead, show a strong positive correlation between LRP and softmax robustness, especially as the number of posterior samples increases. While it is immediately evident that Bayesian predictions are robust to adversarial attacks, it is also clear from this correlation that attacks which are more successful also alter more substantially the interpretation of the classification, and are likely to represent genuine directions of change of the true underlying decision function along the data manifold.

Conclusions Our geometric analysis suggests a fundamental link between the fragility of DNNs against adversarial attacks and the difficulties in understanding their predictions: gradients of the loss function and the prediction function tend to be dominated by directions which are orthogonal to the data manifold. These directions both give rise to adversarial attacks and provide spurious explanations which are orthogonal to the natural parameterisation of the data manifold. We point out the presence of theoretical and practical limitations. The strong assumptions in our Theorem, which restrict the geometrical considerations to fully trained BNNs in the limit of an infinite amount of weights and training data, do not prevent us from observing the desired behavior in practice, even when using cheap approximate inference techniques (VI). However, learning accurate BNNs on more complex datasets is extremely challenging, which makes the Bayesian scheme currently not suitable for large-scale applications. This suggests the need for further investigations on such matters, especially on sufficiently accurate and scalable approximate inference methods for BNNs such applications [15]. Nevertheless, we believe that the insights provided by a geometric interpretation will be helpful towards a better understanding of both the strengths and the weaknesses of deep learning.

References

[1] David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. *arXiv preprint arXiv:1806.07538*, 2018.

[2] Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pages 314–323. PMLR, 2020.

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[4] Kirill Bykov, Marina M-C Höhne, Klaus-Robert Müller, Shinichi Nakajima, and Marius Kloft. How much can i trust you?–quantifying uncertainties in explaining neural networks. *arXiv preprint arXiv:2006.09000*, 2020.

[5] Ginevra Carbone, Matthew Wicker, Luca Laurenti, Andrea Patane, Luca Bortolussi, and Guido Sanguinetti. Robustness of bayesian neural networks to gradient-based attacks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15602–15613. Curran Associates, Inc., 2020.

[6] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

[7] 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.

[8] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

[9] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019.

[10] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[11] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[12] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

[13] Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *stat*, 1050:22, 2018.

[14] Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[15] Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

[16] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. Interpretable deep learning under fire. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020.