Are Neural Networks Robust to Gradient-Based Adversaries Also More Explainable? Evidence from Counterfactuals

Rithik Appachi Senthilkumar [1] (rappachisenthi@tudelft.nl) Patrick Altmeyer [1] Dr. Cynthia Liem [1]

[1]EEMCS, Delft University of Technology

1. Background

Neural networks are vulnerable to **adversarial attacks** – tiny perturbations to input data that elicit misclassifications [4]. Gradient-based adversaries leverage the network's gradients to create perturbations. **Adversarial training** makes neural networks robust to such attacks, by training them on both regular and adversarial examples.

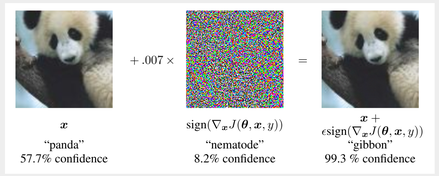

Figure 1. A demonstration of the Fast Gradient Sign Method (FGSM) attack [2], a gradient-based adversarial attack causing a model to misclassify a panda as a gibbon.

Counterfactual Explanations give us insights into how machine learning models make decisions, by exploring how strategically modifying certain feature values causes changes in model outputs.

Figure 2. A counterfactual '9' generated for a factual '3', for a neural network

Plausible counterfactuals are those consistent with the distribution of the data, while faithful counterfactuals are those consistent with the model's *learned representation* of the data.

Model explainability can be defined as the *degree to which faithful counterfactuals generated for the model are also plausible*.

2. Research Questions

Prior work qualitatively demonstrated that adversarially robust models produced counterfactuals containing more class-specific features than regular models, for image data.

We perform a systematic, quantitative study across both tabular and image data that investigates the following questions:

Research Question 1: *Are neural networks made robust to gradient-based adversaries through adversarial training more explainable than a regularly trained neural network?*

Research Question 2: *Among adversarially robust neural networks, are those trained with a stronger gradient-based adversary more explainable than those trained with a weaker gradient-based adversary?*

3. Methodology

We perform experiments for both image data (MNIST) and tabular data (California Housing), to determine whether robust neural networks are more explainable than standard networks, and whether the extent of adversarial training impacts explainability. We perform the following steps:

- Train a regular neural network (no adversary), and three networks with varying strengths of adversary in training using the gradient-based *Projected Gradient Descent (PGD) attack*. We measure their robustness against FGSM [2] and PGD [3] attacks.
- Generate counterfactuals for each of our models both along **inter-class decision boundaries**, and in **the model's learned maximum likelihood regions for the target class**, using the *Energy-Constrained Counterfactuals (ECCo)* [1] generator that produces counterfactuals faithfully describing model behavior.
- Compare the plausibilities of faithful counterfactuals produced by *ECCo* for our regular and robust networks. We measure the implausibility of a counterfactual as the average distance between itself and its nearest neighbors.

4. Results

Table 1. Implausibilities and robust accuracies for standard and robust neural networks trained on MNIST and California Housing. Lowest implausibility for each dataset and convergence criterion marked in bold

Dataset	Training	Accuracies			Counterfactual Implausibilities	
		Clean	FGSM	PGD	Impl. (Boundary)	Impl. (Target Class)
MNIST	Standard	0.981	0.032	0.002	0.437 ± 0.002	0.390 ± 0.004
	Strong-AT	0.969	0.714	0.653	0.412 ± 0.002	$\mathbf{0.221 \pm 0.005}$
	Medium-AT	0.984	0.379	0.298	$\mathbf{0.411 \pm 0.002}$	0.236 ± 0.005
	Weak-AT	0.983	0.248	0.062	0.416 ± 0.003	0.262 ± 0.006
California Housing	Standard	0.857	0.211	0.217	1.673 ± 0.072	$\mathbf{2.358 \pm 0.124}$
	Strong-AT	0.771	0.644	0.647	$\mathbf{1.196 \pm 0.072}$	2.762 ± 0.149
	Medium-AT	0.810	0.563	0.572	1.224 ± 0.070	2.884 ± 0.188
	Weak-AT	0.840	0.337	0.350	1.390 ± 0.074	2.632 ± 0.122

Our key takeaways as observed in Table 1 are as follows:

- Robust neural networks for both datasets learned more explainable decision boundaries between classes than standard networks. For California Housing data, more robust networks produced more plausible counterfactuals.
- Robust neural networks for image data (MNIST) learned more explainable representations of classes than standard networks, with more robust models producing more plausible counterfactuals.
- Surprisingly, robust California Housing networks did not learn more explainable class representations than the regular network, most likely because adversarial training worsened the model's learned class representations to accommodate for adversarial examples. This is evidenced by the high clean and robust accuracy trade-off.
- The above underscores difficulties with traditional adversarial training for tabular data, due to the innate properties of tabular data (heterogeneity, varying inter-feature correlations).

5. Visual Example

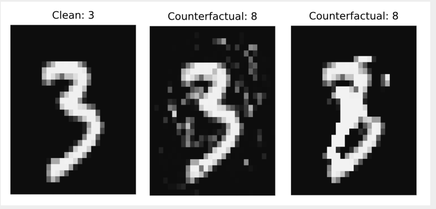

Figure 3. A factual '3' (left) and counterfactual '8's produced by the standard model (middle) and most robust model (right)

This example demonstrates a robust MNIST neural network producing a counterfactual significantly more plausible than that of a standard network.

6. Limitations and Future Work

- Our research was limited to the scope of gradient-based adversaries. However, there also exist **black-box adversaries** that learn about the underlying model by querying it. Future work can explore how black-box robustness impacts model explainability.
- We used one network architecture for each of our datasets, which may not comprehensively capture the relationship between robustness and explainability.
- Our experiments revealed difficulties with adversarial training techniques for tabular data. An interesting direction for future research can be using deep neural decision tree architectures for tabular data, which perform better than shallow neural networks.

7. Conclusion

- We demonstrated empirically that neural networks robust to gradient-based adversaries learned more explainable boundaries between classes, for both image and tabular data
- For image data, we observe that robust networks learn significantly more explainable class representations than a standard network, and that the more robust a neural network is, the more explainable its learned representation of the classes.
- We hope our research encourages future work towards developing robust neural networks paying adequate consideration to model explainability.

References

[1] Patrick Altmeyer, Mojtaba Farmanbar, Arie van Deursen, and Cynthia CS Liem. Faithful model explanations through energy-constrained conformal counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10829–10837, 2024.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[3] 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.

[4] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.