Verified Explanations of Misclassified and Adversarial Inputs

Xiaofu Li Mentored by Clark Barrett and Min Wu

Background

VeriX

As modern day neural networks become more complicated in architecture, more information gets folded in the deep layers and they become more and more of "black-box" processes. In order for us to trust decisions made using neural networks, there emerges the need to produce explanations for network predictions.

Compared to heuristic explanation methods, VeriX provides optimal robust explanations. Given a perturbation ϵ, VeriX divides all input features into a relevant and an irrelevant set with the following guarantees:

1. *Robust*: perturbation on features in the irrelevant set within ϵ will never change the prediction of the network.
2. *Optimal*: no feature in the relevant set can be moved to the irrelevant set without violating the first guarantee.

Adversarial samples and adversarial training

While neural networks are very powerful at many tasks such as image classification, it is possible to craft adversarial inputs that are close to the original input so that they are virtually indistinguishable to human eyes but can fool neural networks to label them as something else. This poses a security risk where malicious data can disguise themselves as benign.

Methods to counter adversarial inputs can usually be classified as detecting them or making networks more robust. A method to make networks more robust is to mix in adversarial samples during the training of networks, called adversarial training.

Project Goals

In this project, I aim to compare explanations generated by VeriX on different types of networks and inputs. We want to answer the following questions:

- How do explanations for **correct and incorrect predictions** differ?
- How do explanations for predictions made on **normal versus adversarial samples** differ?
- How does **adversarial training on networks** affect explanations generated?

Methods

I used the MNIST dataset and trained two two-layer fully connected networks with 10 units in each layer, one regularly trained, one adversarially trained. For adversarial training, I used projected gradient descent (PGD) attack to generate adversarial samples on half of the training data after each epoch and mixed them with the rest of training data.

Adversarial test data are generated on a different network trained on the MNIST dataset also using PGD attack. The regular 10x2 network achieves 93.95% accuracy on original test data but only 25.82% accuracy on adversarial samples. The adversarially trained 10x2 network achieves 92.85% accuracy on original test data and 82.66% accuracy on adversarial samples.

I first compared VeriX explanations for normal versus adversarial inputs and normal network versus adversarial network using 100 test samples. Then, for both networks with both real and malicious samples, I took 100 correct predictions and 100 incorrect predictions each to generate explanations for. For all experiments, ϵ value is set to 0.05, and explanation size is measured by pixel numbers.

Results

Inputs with adversarial perturbation applied tend to have a larger explanation size compared to the original inputs. Among 100 test samples, 93 had a larger explanation size after adversarial perturbation, with an average difference of 383.33 pixels.

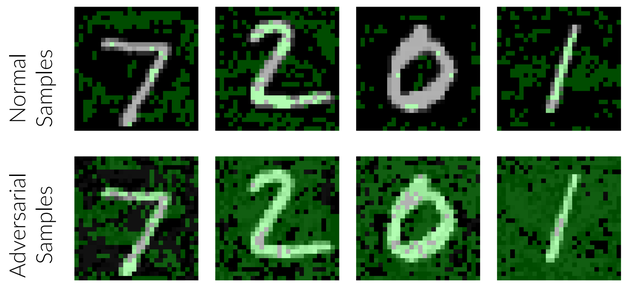

Figure 1. Comparison of explanations for normal samples and adversarial samples on the normal 10x2 network.

Adversarially trained network tends to have a smaller explanation size than the normal network. It is also much more likely to achieve ϵ-robustness, meaning perturbation for all pixels within ϵ won't change the prediction. In 100 test samples, the adverarial network achieved ϵ-robustness when $\epsilon = 0.05$ on 84 samples, in comparison to 2 samples for the normal network. This means small perturbations is less likely to change the predictions of the adversarial network. Figure 2 shows examples where the relevant set of explanation generated on the adversarial network is not empty.

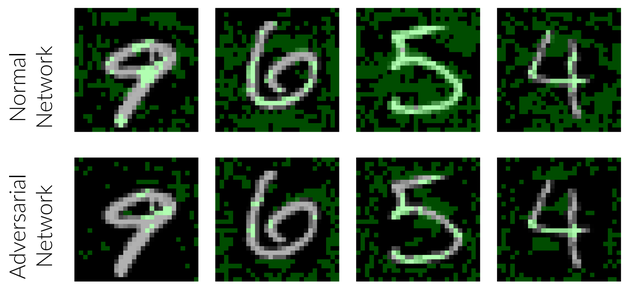

Figure 2. Comparison of explanations produced by normally trained and adversarially trained 10x2 networks.

Table below summarizes results from experiments with correct and incorrect predictions on different networks and input types.

	Normal Network		Adversarial Network	
	Correct	Incorrect	Correct	Incorrect
Normal samples	177.97	399.84	44.68	319.42
Adversarial samples	524.9	556.72	268.63	549.92

Table 1. Average explanation sizes for 10x2 networks on MNIST. 100 samples are selected for each category.

To check if the explanation size differences came from different levels of confidence in prediction or properties of the networks and inputs, I also plotted the relationship between explanation sizes and pre-softmax logit values of predictions.

For the normal network on real input samples, explanation size is correlated to the predictions logit values. Both act as predictors of correct or incorrect classification.

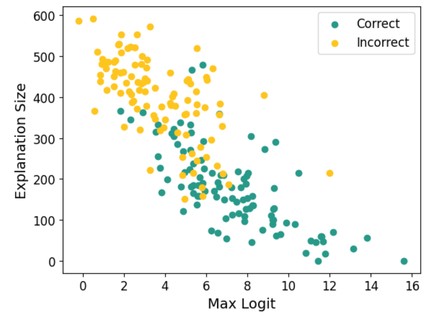

Figure 3. Explanation size versus prediction logit value for correct and incorrect predictions made by the normal network on real inputs.

For adversarial samples and real samples, logit values show similar distributions, but adversarial samples tend to have larger explanation sizes even when logits are similar.

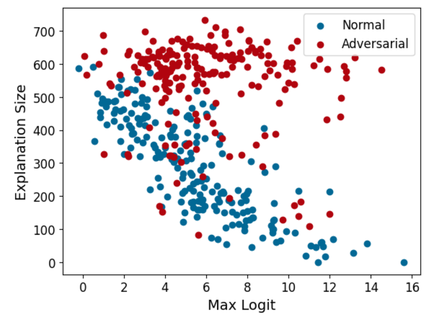

Figure 4. Explanation size versus prediction logit value for real and adversarial samples passed through the normal network, including 100 correct and 100 incorrect predictions for both real and adversarial samples.

Conclusion

- Incorrect predictions have larger explanation sizes than correct ones
- Adversarial samples have larger explanation sizes than real samples even for predictions with similar confidence.
- Adversarially trained networks result in smaller explanation sizes compared to normal networks.

Future Work

- Perform confidence tests to evaluate how good a predictor explanation size is for incorrect classifications and adversarial inputs.
- Run experiments on different datasets to see if similar effects are observed.
- Test different network architectures, including bigger fully connected networks and convolutional neural networks.

References

[1] Min Wu, Haoze Wu, and Clark Barrett.
Verix: Towards verified explainability of deep neural networks, 2023.