May 7th, 11:00 AM - 1:00 PM

Reliable Explanations via Adversarial Examples on Robust Networks

Walt Woods
Portland State University

Jack H. Chen
Portland State University

Christof Teuscher
Portland State University

Let us know how access to this document benefits you.
Follow this and additional works at: https://pdxscholar.library.pdx.edu/studentsymposium

Part of the Electrical and Computer Engineering Commons

Woods, Walt; Chen, Jack H.; and Teuscher, Christof, "Reliable Explanations via Adversarial Examples on Robust Networks" (2019). Student Research Symposium. 28.
https://pdxscholar.library.pdx.edu/studentsymposium/2019/Posters/28

This Event is brought to you for free and open access. It has been accepted for inclusion in Student Research Symposium by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.
Reliable Explanations via Adversarial Examples on Robust Networks

Walt Woods, Jack Chen, and Christof Teuscher; teuscher:Lab, Department of Electrical and Computer Engineering, Portland State University, Portland, OR, USA; {wwoods, chenjac, teuscher}@pdx.edu

Abstract

Neural Networks (NNs) are increasingly used as the basis of advanced machine learning techniques in sensitive fields such as autonomous vehicles and medical imaging. However, NNs have been found vulnerable to a class of imperceptible attacks, called adversarial examples, which arbitrarily alter the output of the network. We propose a new method for stabilizing networks, and show that as an added bonus, our technique results in reliable, high-fidelity explanations for the NN’s decision.

Methods + Results

Creating a classifier which is robust to adversarial perturbations is the same as solving a saddle-point formulation for finding the best-case classifier when presented with the worst-case input perturbation:

$$\min_{\mathbb{E}[x,t]} \max_{\theta \in S} L(\theta, x + \epsilon \cdot t)$$

Madry et al. showed that the solution to this equation could be approximated by sampling the inner maximum [1]. Later work by Tsipras et al. demonstrated that this method of training could result in salient features being produced in adversarial examples [2].

We extended this with a variation of the Lipschitz constraint, which dictates that the derivative of a continuous function shall not exceed some bound:

$$|f(x_1) - f(x_2)| < K|x_1 - x_2|$$

In effect, this limits the efficacy of the maximum in the saddle formulation. This, along with other modifications, led to networks against which adversarial attacks demonstrate salient features.

Citations:

Conclusions

- Adversarial examples may be used to explain NNs by demonstrating the nearest decision boundary. A form of Lipschitz continuity can be used to dramatically improve the quality of these adversarial examples.
- The proposed modifications make it possible to generate reasonable images of a target class based on a source image.
- Paper pending.