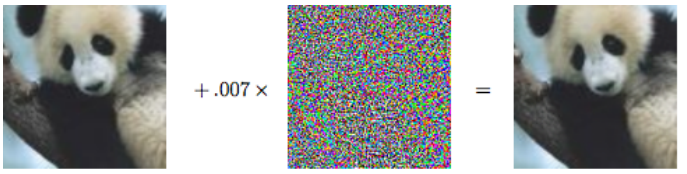
A Comprehensive Benchmark for Defenses Against Black-Box Adversarial Attacks

Background



[1]

Adversarial Examples are a fundamental problem observed in Deep Learning, and potentially a huge safety and security issue. Tiny perturbations of an input can lead to completely undesired behaviour of systems that were believed to be accurate. But how big is this threat in reality? What does it mean for Machine Learning theory? And what can we do about it?

Our team is based at the Chair of Robotics, Artificial Intelligence and Real-time Systems at TUM, and at fortiss, which is the the new Bavarian research center for AI. Our work encompasses safety and security for both the practical application of AI systems (e.g. autonomous driving), and also the very fundamentals of Machine Learning. Recently, we won a prize in the prestigious NeurIPS 2018 Adversarial Vision Challenge for our work on black-box adversarial attacks - fooling Neural Networks is a lot of fun!

Thesis topic

Recently, a large number of papers has been published that claim robustness against black-box adversarial attacks. This stands in direct contrast to our own findings, and arguably the state of the art: adversarial robustness currently does not exist, and most defenses use cheap tricks that are easily circumvented. Many peer-reviewed(!) publications claim strong defensive capabilities, while at the same time proven attacks exist that contradict their statements. We suspect this problem has arisen due to the lack of an established benchmark in a black-box scenario.

The main goal of this study is to provide a comprehensive analysis of existing black-box defenses, and to derive from it the true state of the art in black-box robustness. For this, the proposed defenses are to be reimplemented, along with our state-of-the-art attacks, and evaluated in meaningful scenarios. This work will also result in recommendations for a new benchmark, against which black-box Adversarial Examples can be evaluated in the future.

Requirements

- Basic knowledge of Adversarial Examples
- Excellent programming skills
- Proficiency in at least one Deep Learning framework (PyTorch, TensorFlow, ...)
- Demonstrated experience with Neural Networks

Please send your CV via e-mail, along with a transcript of records and meaningful references. Your Machine Learning experience should be clearly visible. Code (GitHub etc.) is a huge plus.

Disclaimer: Only students at TUM can be supervised. This document is not a specific thesis offer, but a general description of an area of research.

[1] Goodfellow, I., Shlens, J., and Szegedy, C., "Explaining and Harnessing Adversarial Examples". In *International Conference on Learning Representations*, 2015.

fortiss

Research Institute of the Free State of Bavaria

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems

Supervisor:
Prof. Dr.-Ing. Alois Knoll

Advisor:
Thomas Brunner

Type:
Master's Thesis

Research area:
Artificial Intelligence

Language:
English

For more information please contact us:

E-Mail: brunner@fortiss.org

Web: www.fortiss.org