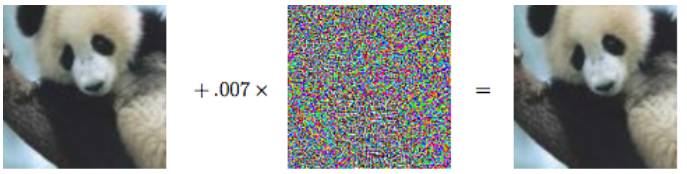
Cognition-based Metrics for Adversarial Robustness

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

How does one measure robustness? Currently, virtually all researchers in the field of Adversarial Examples use metrics such as ℓ_∞ or ℓ_2 (euclidean) distance. These measures are easy to optimize, but they are not adequate for assessing safety at all. Arguably, the main threat of Adversarial Examples is that they fool the machine, but not the human. Therefore, we must measure robustness in terms of human cognition - and evaluate Adversarial Examples against that standard.

There exist several metrics in other fields that are loosely based on human cognition - or so they claim. This work will evaluate these metrics in the context of Adversarial Examples, discuss their usefulness, and give valuable pointers for future research in the field. If successful, this work has the potential to make a big impact.

Requirements

- 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