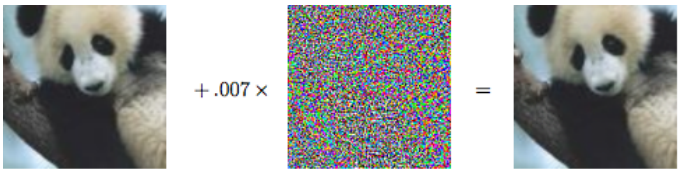
Using Reinforcement Learning to Find Adversarial Examples

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

Finding Adversarial Examples usually involves an iterative search procedure. Attacks traverse the input space in an agent-like fashion, at each step deciding which direction to take. Little is known about this complicated landscape, and black-box attacks often need to employ expensive random search techniques. How can we increase the efficiency of that search? Framed in this way, the problem seems like a perfect candidate for Reinforcement Learning. Could we train an agent that learns to choose a good adversarial perturbation? How much is the speedup? Does it transfer across different architectures?

This work combines two exciting topics, Adversarial Examples and Reinforcement Learning. If successful, it will yield more powerful adversarial attacks than ever before, and also allow us to gain key insights into the regions in which Adversarial Examples exist. This project is fairly ambitious - please apply only if you understood this description.

Requirements

- Excellent programming skills
- Proficiency in at least one Deep Learning framework (PyTorch, TensorFlow, ...)
- Demonstrated experience with either Adversarial Examples or Reinforcement Learning. There will not be enough time to learn both.

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