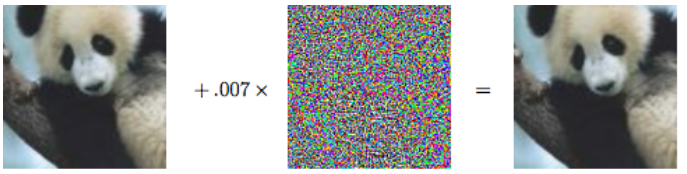
Adversarial Examples That Fool Classic Computer Vision Methods

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

There exists the widespread notion that classic computer vision is somehow "safe", while Deep Learning is unsafe by design. But is that really true? So far, research on Adversarial Examples has focused on Neural Networks exclusively. But with the advent of strong black-box attacks - that work without any information about the model - it has become possible to attack virtually any classifier.

This work aims to pit state-of-the-art adversarial attacks against classic computer vision methods in a range of established scenarios and determine just how robust these "tested and true" methods really are. This will identify some of the key factors in robustness - both on the algorithm and use case level - and help the ongoing AI safety debate with some much-needed fresh air.

Requirements

- Excellent programming skills
- Proficiency in at least one Deep Learning framework (PyTorch, TensorFlow, ...)
- Demonstrated experience in computer vision (HOG, SIFT, Fisher Vectors, ...)

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:
BA, MA

Research area:
Artificial Intelligence

Language:
English

For more information please contact us:

E-Mail: brunner@fortiss.org

Web: www.fortiss.org