Naturalistic Adversarial Attack Algorithms on Deep Neural Networks

Adversarial vulnerability of deep neural networks (DNNs) is a well-studied phenomenon, where minimally perturbed input examples by an attacker can unreasonably alter the usual inference process of the network. More broadly however, DNNs are also known to be vulnerable to simple distribution shifts. At test time, unlike human visual perception, naturalistic changes in the image content (e.g., brightness, fog, or snow as illustrated above) can also cause the input images to lie far from the in-distribution density introduced by the training set images, hence leading to a significant performance loss in terms of generalization. In this project we will aim to align these two aspects by developing algorithms to craft naturalistic adversarial attacks for state-of-the-art deep learning models. We will develop novel adversarial attack algorithms and evaluate their effectiveness with respect to state-of-the-art defenses on different deep learning architectures.

Goals & Tasks

- Review of the state-of-the-art on adversarial attacks and out-of-distribution generalization in deep learning.
- Development and implementation of naturalistic adversarial attack algorithms.
- Experimentally evaluating attacks on existing defenses and architectures.

Qualifications

- Interest in deep learning.
- Experience with Python based deep learning frameworks such as TensorFlow or PyTorch are beneficial.
- Registered to one of the following:
  - ✓ **Bachelor Thesis**
  - ✓ **Seminar Project**
  - ☐ **Master Thesis**

  Extension for a Master Thesis is possible on interesting results.

Contact

Ozan Özdenizci
ozan.ozdenizci@igi.tugraz.at