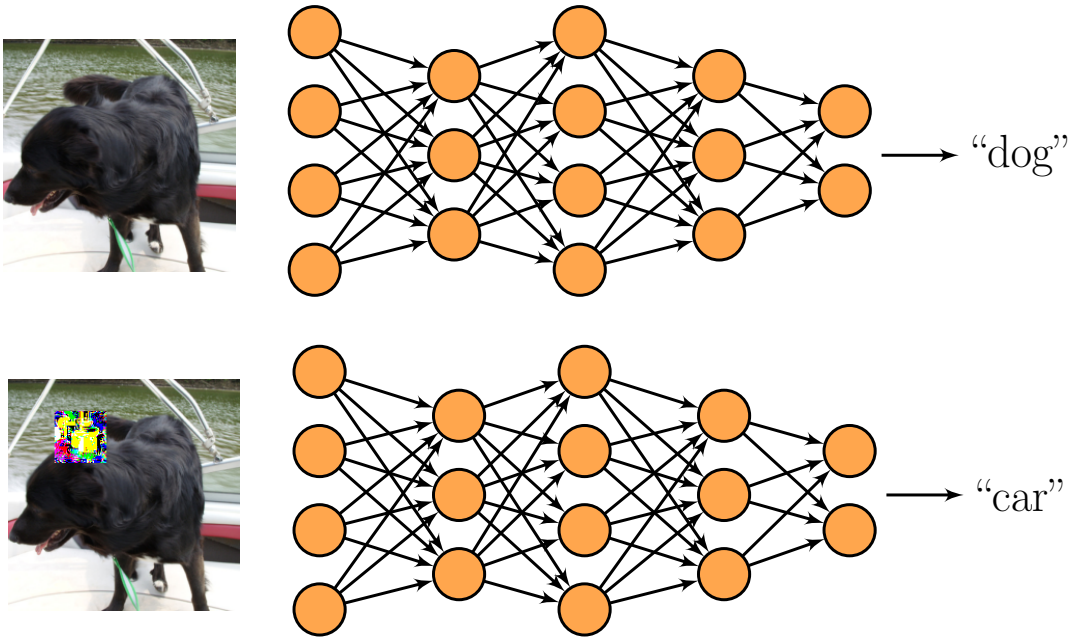
Exploring Real World Vulnerability of Deep Neural Networks Against Adversarial Patch Attacks



Adversarial vulnerability of deep neural networks (DNNs) is a well-studied phenomenon, where minimally perturbed input examples can unreasonably alter the usual inference process of the network. This concept extended further to physical-world adversarial attacks in less than a decade ago, where one can create a physically realizable, adversarially patterned patch for malicious purposes against DNNs (e.g., printed adversarial stickers that can be physically placed on objects). In this project we want to explore different aspects of this machine learning security concern in terms of generalization and real-world reliability across different optimization algorithms and DNN architecture types.

Goals & Tasks

- Review and implementation of state-of-the-art on adversarial patch attacks.
- Evaluating alternative defenses and DNN training algorithms under attacks.
- Simulating real-world reliability of adversarial patch attacks and defenses.

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**

Contact

Ozan Özdenizci
ozan.ozdenizci@igi.tugraz.at