Adversarial Attacks and Defences

Date: 13/10/2023

Team ID: Team591481

Project name: Advesarial attacks and defences

Team members: Debjit Mondal

Abstract:

Hello, this is Bing. I can try to write an abstract for this project based on the information available on GitHub. Here is my attempt:

Adversarial Attacks and Defences is a project that aims to provide a comprehensive survey of the state-of-the-art techniques and challenges in the field of adversarial machine learning. The project consists of a paper and a code repository that demonstrate various types of adversarial attacks and defences on deep learning models.

Adversarial machine learning is a branch of machine learning that studies the vulnerability of deep learning systems to malicious inputs, known as adversarial examples, that can fool the model into making incorrect predictions. Adversarial examples pose a serious threat to the security and reliability of many applications that rely on deep learning, such as computer vision, natural language processing, speech recognition, and biometrics.

The project paper covers the following topics:

- The motivation and background of adversarial machine learning, including the definition, formulation, and properties of adversarial examples.
- The taxonomy and classification of different types of adversarial attacks, such as evasion and poisoning attacks, and white-box and black-box attacks, along with their examples and applications.
- The taxonomy and classification of different types of adversarial defences, such as adversarial training, gradient masking, detection, and certification, along with their advantages and limitations.
- The evaluation and comparison of the effectiveness and robustness of various adversarial attacks and defences on benchmark datasets and models, such as MNIST, CIFAR-10, ImageNet, and ResNet.

- The open challenges and future directions of adversarial machine learning, such as transferability, scalability, interpretability, and generality.

The project code repository contains the implementation of several adversarial attacks and defences in Python using TensorFlow and Keras frameworks. The code is modular and easy to use, and can be applied to any deep learning model and dataset. The code also provides the visualization and analysis of the results, such as the accuracy, loss, and confusion matrix of the model under different attack and defence scenarios.

The project is intended to serve as a useful resource and reference for researchers, practitioners, and students who are interested in learning more about the fascinating and emerging field of adversarial machine learning. The project is open-source and welcomes contributions from the community.