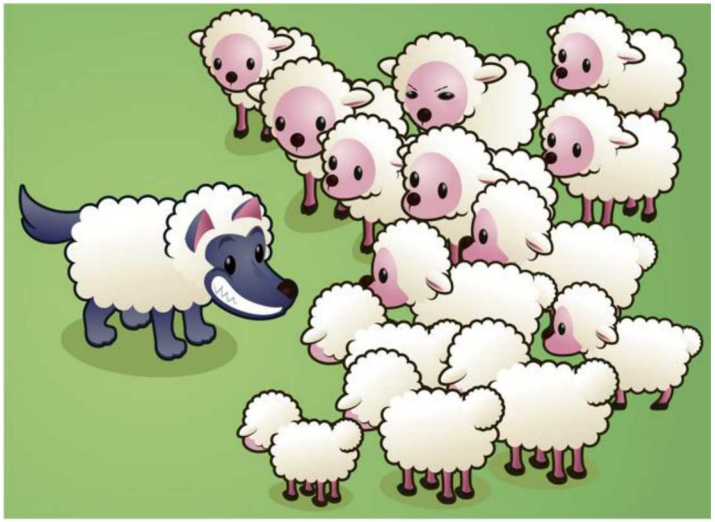
SAND2018-8252PE

A Brief Survey of Adversarial Concerns in Machine Learning and Deep Learning



Philip Kegelmeyer, Sandia/CA, 8700

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

July 31, 2018

A Terse List of Adversary Goals

Subvert the Model:

Adjust the training data to undermine the model.

Evade the Model:

Adjust the test data to avoid correct classification.

Reveal the Model:

Cause the model to reveal more than was intended.

A Terse Glossary

Machine Learning (ML):

Converting useful features into classifications.

Feature Learning (FL):

Converting raw data into useful features.

Deep Learning (DL):

Feature learning followed by machine learning.

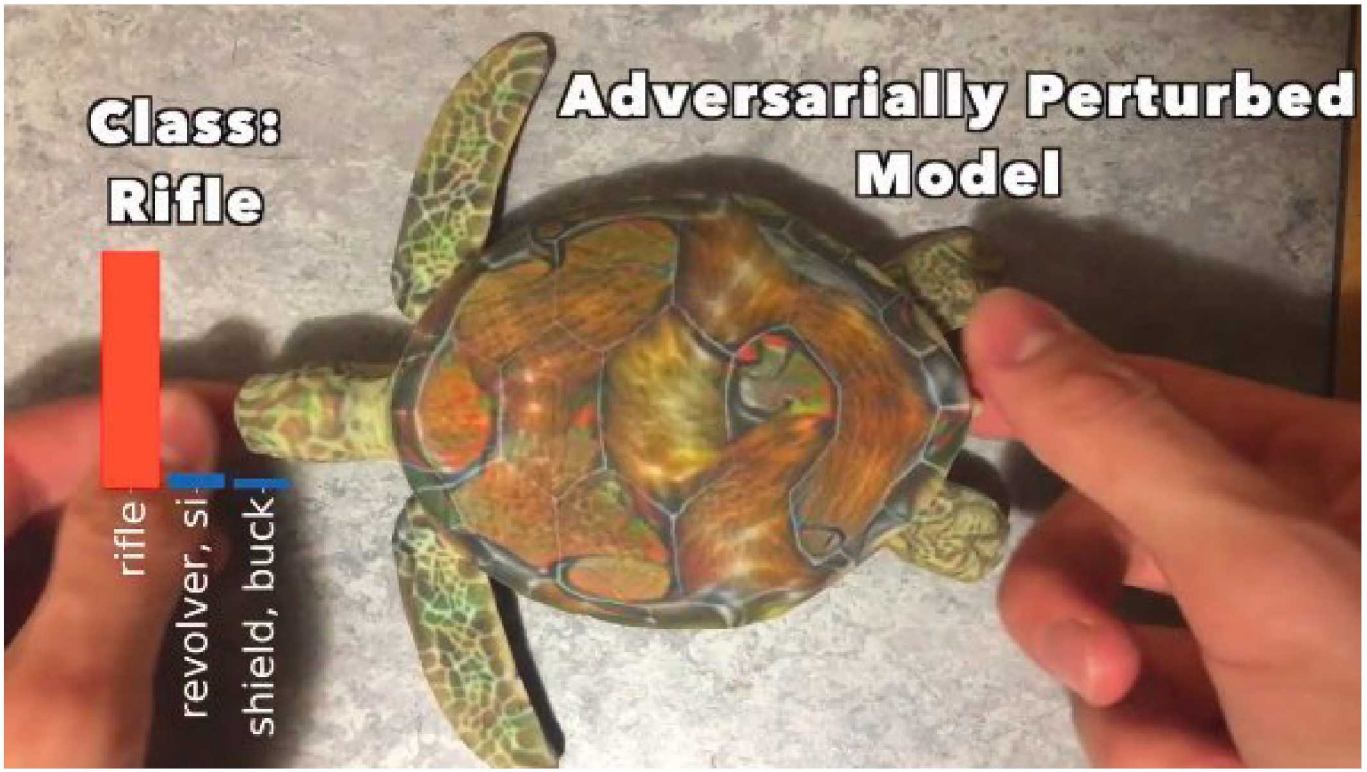
Transfer Learning (TL):

Using a pre-trained network to jumpstart feature learning.

Test Sample Attacks Deep Learning Image Analysis

Subvert Evade Reveal / ML DL→ML TL→DL→ML

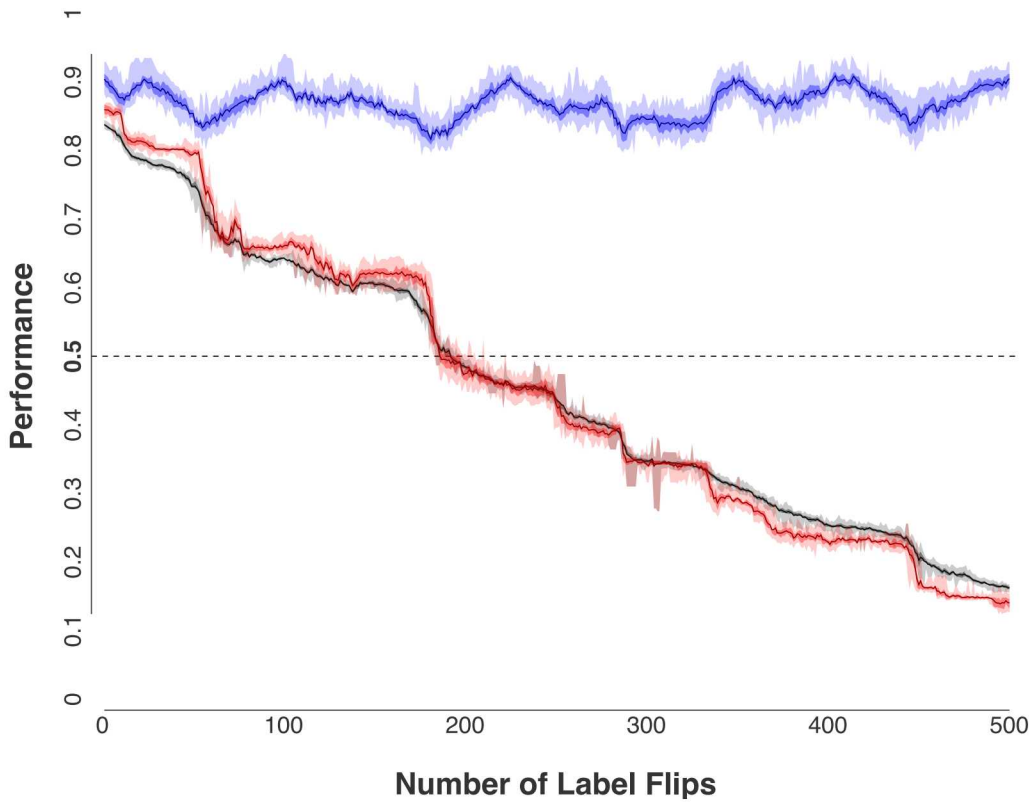
Many recent examples.



Synthesizing Robust Adversarial Examples, Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok

Label Tampering to Reduce Accuracy (Sandia)

Subvert Evade Reveal / ML DL→ML TL→DL→ML



Brute Clustering

cross-validation on training, **ensemble performance on test**, **non-ensemble performance on test.**

Model Stealing

Subvert Evade Reveal / ML DL→ML TL→DL→ML

Copying or reverse engineering a machine learning model.

Credit: Dooder, Freepik.com

Insert Trojans into Training Data

Subvert Evade Reveal / ML DL→ML TL→DL→ML

Backdoors inserted into DL models through small changes to training data.

Credit: *BadNets: ...*, Gu et. al.

These backdoors can persist even after alterations due to transfer learning.

Model Inversion and Memorization

Subvert Evade Reveal / ML DL→ML TL→DL→ML

Subvert Evade Reveal / ML DL→ML TL→DL→ML

Machine learning models can be subverted or probed to:

- Recover training data that was inadvertently memorized (specific credit card numbers in text)
- Estimate whether a given sample was used to train the model
- Recreate data used to train the model

training image

recovered from model

Pointers to Papers, Part 1

- Label Tampering to Reduce Accuracy
  - *Counter Adversarial Data Analytics*, Philip Kegelmeyer et. al, Sandia Report, SAND 2015-3711, May 2015.
- Deep Learning Image Analysis Test Sample Attacks
  - *Robust Physical-World Attacks on Machine Learning Models*, Kevin Eykholt, et. al.
  - *Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples*, Nicolas Papernot, et. al.
  - *Synthesizing Robust Adversarial Examples*, Anish Athalye, et. al.
- Model Stealing
  - *Stealing machine learning models via prediction APIs*, Tramer et. al.
  - *Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples*, Nicolas Papernot, et. al.

Pointers to Papers, Part 2

- Insert Trojans into Training Data
  - *BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain*, Tianyu Gu, et. al.
- Model Inversion and Memorization
  - *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, Matt Fredrikson et. al.
  - *Machine Learning Models that Remember Too Much*, Congzheng Song et. al.
  - *The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets*, Nicholas Carlini
  - *Membership Inference Attacks against Machine Learning Models*, Reza Shokri et. al.