Detecting Adversarial Examples for Machine Learning Models

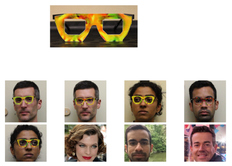
IoBT REIGN

Relevant to IOBT CRA Research Area / Task: Research Area 2, Task 2.3

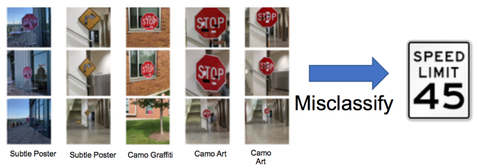
Objective

Detecting Adversarial Attacks on Machine Learning Models

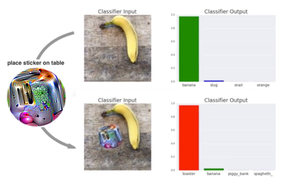
Szegedy et al, 2013, 2014



Sharif et al, 2016

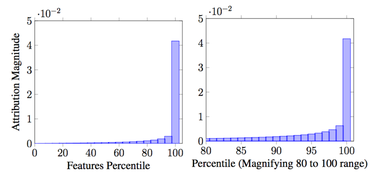


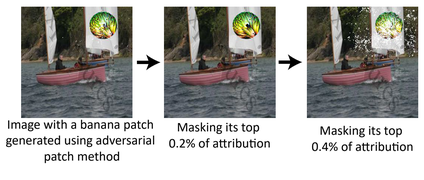
Evtimov et al., 2017

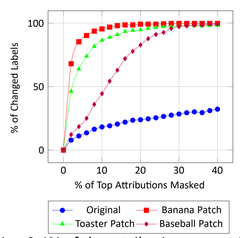


Brown et al. 2017

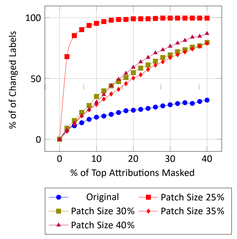
Results







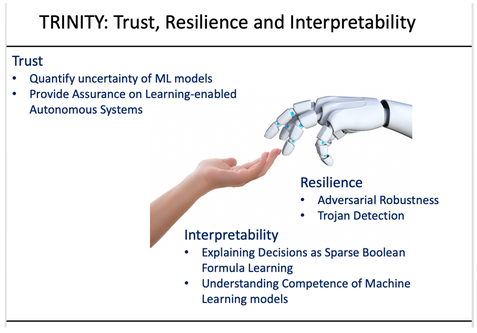
Dropping 0.4% of the attribution causes 99.71% of the attacks based on banana patches, 98.14% of the attacks based on toaster patches, and 99.20% of the attacks based on baseball patches to be detected.

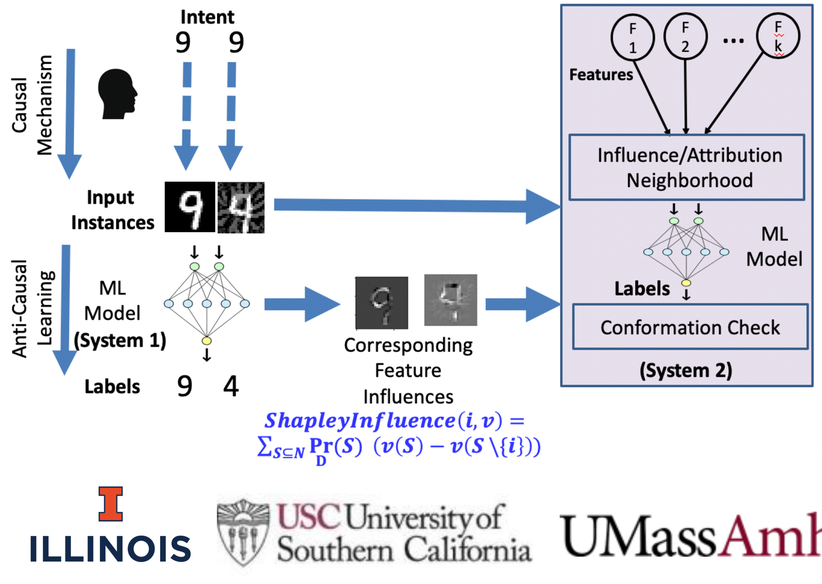


Masking 0.4% of attributions caused nearly 80% of labels to change for images with adversarial patches.

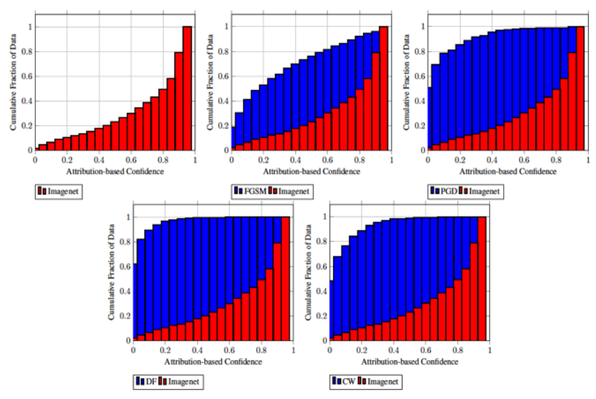
Approach

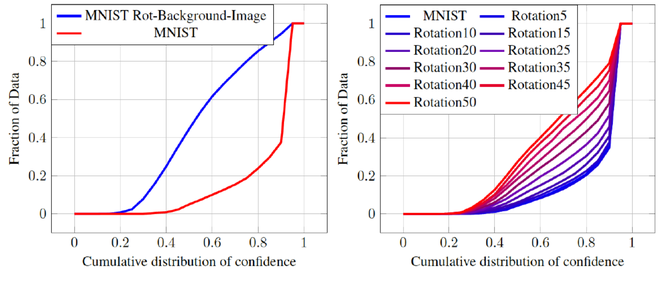
Compute attribution of features
Construct causal neighborhood using attribution for importance sampling
Measure conformance of ML model in this neighborhood.





$$ShapleyInfluence(i, v) = \sum_{S \subseteq N} \Pr_{D}(S)\ (v(S) - v(S \setminus \{i\}))$$





Conclusions

- We employ an attribution-driven sampling of the neighborhood of a given input and measure the conformance of the model's predictions to compute the attribution-based confidence (ABC) metric for DNN prediction on this input.
- While directly sampling the neighborhood of a high-dimensional input is challenging, our approach uses attribution-based dimensionality reduction for finding locally relevant features in the vicinity of the input, which enables effective sampling.
- We theoretically motivate the proposed ABC metric from the axioms of Shapley values, and experimentally evaluate its utility over out-of-distribution data and adversarial examples.
- Our approach is particularly suitable for detecting adversarial patch attacks which are physically realizable.
- Smaller the patch, the better our approach will perform in detecting the adversarial attack!

Path Forward

- Replace conformance with more fine-grained measures such as distribution shift metrics.
- Temporal evolution of attributions to detect adversarial attack at instance of occurrence.

- Jha et. al. Attribution-Based Confidence Metric For Deep Neural Networks, NeurIPS'19
- Jang et. al. On the Need for Topology-Aware Generative Models for Manifold-Based Defenses, ICLR'20
- Jha et. al. Detecting Adversarial Examples Using Data Manifolds, MILCOM'18
- Kiourti et. al. TrojDRL: Trojan Attacks on Deep Reinforcement Learning Agents, DAC' 20

ILLINOIS · USC University of Southern California · UMass Amherst · UCLA · SRI International · Georgetown University

POINT OF CONTACT:

Susmit Jha
susmit.jha@sri.com

Brian Jalaian
brian.a.jalaian.civ@mail.mil