SaTC: CORE: Medium: Collaborative: Towards Robust Machine Learning Systems

Hao Chen (University of California, Davis)

Neil Zhenqiang Gong (Duke University)

Project page: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1801751

Project Overview

- ML Model Robustness
  - Characterize normal and adversarial examples
    - Explore different attack spaces
    - Study the characteristics of the neighborhood
  - Design robust defenses
    - Study the relationship between data and robustness
    - Leverage multiple models with diverse adversarial sensitivity
  - Define metrics for model robustness evaluation
    - Perturbation lower-bounds
    - Computational cost
    - Human perception comparison

Key Challenges

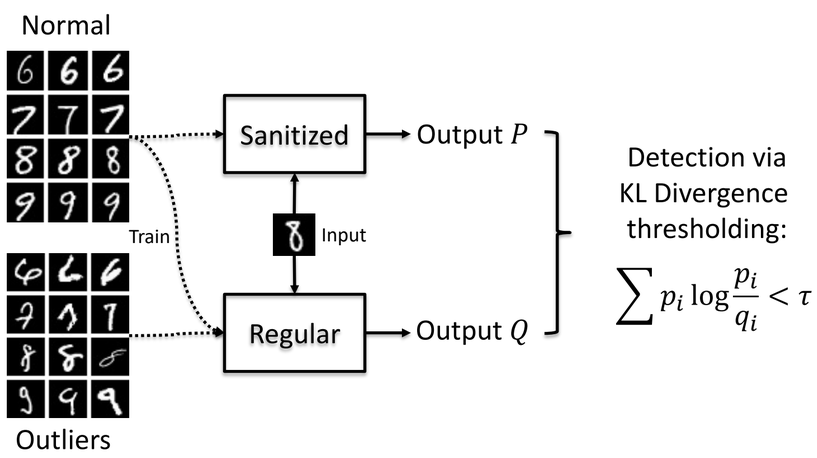
- **Lack of fundamental understanding of how machine learning systems, especially deep learning models, learn information from data and the internal reasonings for each decision made.** A better understanding of these problems can help improve robustness in every stage of building a machine learning system, including data collection, architecture selection, model training, and offline / online evaluations.
- **Hard to measure the boundary of machine perception and the normality of input data in a way comparable with human understanding given a specific task.** To evaluate the real-world robustness risks, instead of studying restricted threat models, researchers need to consider the working boundary of the target system, both in general and ad hoc, as one common goal of various machine learning tasks is to minimize the gap between machine and human perception.

Scientific Impact

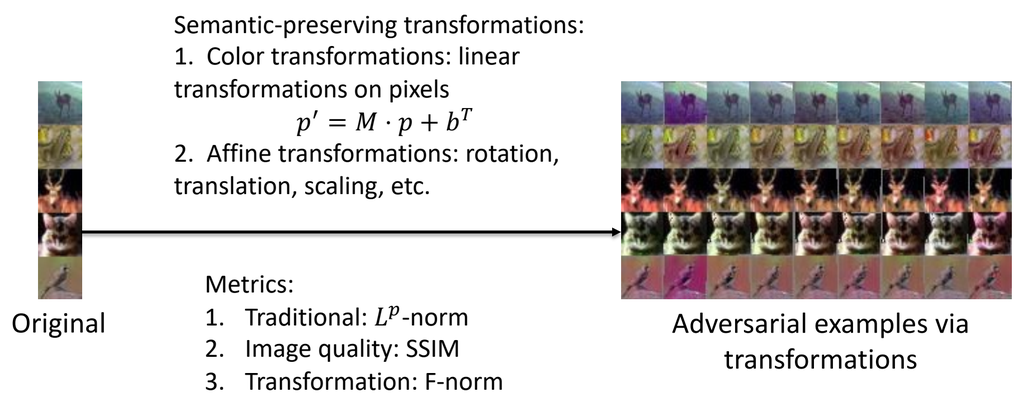
The research of model robustness is highly related to various machine learning research areas. Robustness is an indicator of the worst-case performance of any machine learning system and cannot be neglected. On the other hand, addressing the key challenges can also facilitate related research topics. For example, a better fundamental understanding of machine learning models and their intrinsic vulnerabilities can help develop interpretable AI; studying the contribution of biases in the dataset can provide insights into privacy study in machine learning systems.

Solutions

- **Study the relationship between dataset quality and model robustness**: the study reveals the difference in contributions of training examples provided to the model's decision boundary. We proposed a simple yet effective detection framework by leveraging the finding that outliers in the training dataset may increase the adversarial vulnerability.



- **Explore larger attack spaces in real-world scenarios**: considering semantic-preserving transformations as adversarial attack methods instead of only looking for perturbations in norm restricted spaces. We found that with a larger attack space commonly found in reality, defending against adversarial attacks becomes much harder than ever.



Industrial Impact

Machine learning has been widely applied in the industry. It is undesirable for security-critical tasks such as autonomous driving or intrusion detection to have the robustness of machine learning systems becoming the short board of the bucket. Our project aims to show insights into designing a more proper assessment for real-world adversarial threats and provide suggestions for building systems with better performance-robustness trade-offs.

Educational Impact

Our team will make datasets and source code publicly available and use them in courses and research with graduate and undergraduate students, with particular efforts to include students from underrepresented groups in Science, Technology, Engineering, and Math. The project will also support high school outreach programs and summer camps to attract younger students to study machine learning, security, and computer science.