SaTC: CORE: Small: Adversarial Learning via Modeling Interpretation

PIs: Xia "*Ben*" Hu[1], Guofei Gu[2], James Caverlee[2]
[1]Rice University, Houston, TX 77005 [2]Texas A&M University, College Station, TX 77843

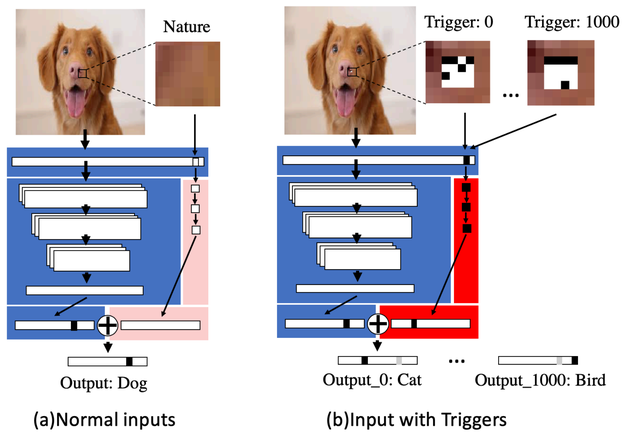
Introduction

AI models are increasingly important in society, with applications including speech recognition, online content filtering, and self-driving cars. However, these models are vulnerable to adversaries attacking them by submitting incorrect or manipulated data with the goal of causing errors, causing potential harm to both the decisions the models make and people who rely on them. The overall project goal is to test this insight and contribute to both the security and data mining communities by developing an adversarial learning framework that leverages interpretability of ML models and results to both identify and mitigate the risks of adversarial attacks.

Scientific Impact

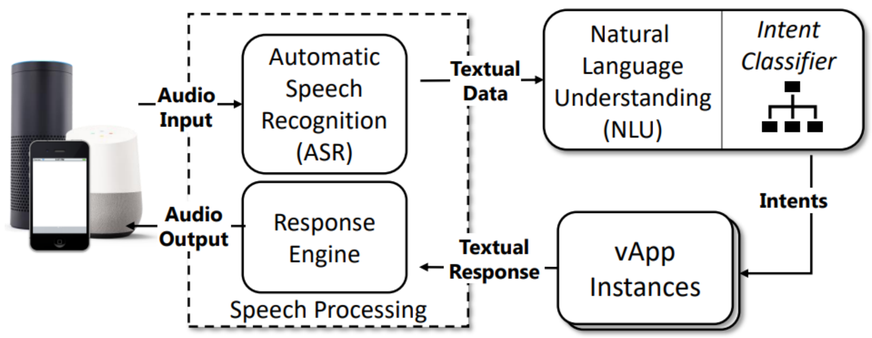
- Develop effective attacking strategies by analyzing modeling interpretation from three aspects including instance level, class level, and a specific group of deep neural networks.
- Explore various defensive strategies to improve the robustness of ML models against adversarial attacks.
- Investigate adversarial learning algorithms to deal with challenges and take advantage of opportunities brought by big data. Specifically, the developed adversarial attacking and defensive algorithms will deal with large-scale, heterogeneous, and relational data.

Technique Approaches



(a) Normal Inputs (b) Inputs with Triggers

We investigated a novel security problem called trojan attack, which aims to attack deployed DNN systems relying on the hidden trigger patterns inserted by malicious hackers [1].



(a) VUI-based VA Architecture

We designed a linguistic-model guided fuzzing tool, named LipFuzzer, to assess the security of Intent Classifier and discover potential misinterpretation-prone spoken errors based on voice command templates [2].

(a) True Input (b) Deepfake (c) LAE

We proposed an active learning framework for deepfake detection, called LAE, which makes predictions relying on correct evidence in order to boost generalization accuracy [3].

Impact on society

This project reveals the severe security problem widely existing in various AI models, which leads to an untrustworthy model towards people. The results remind researchers to revisit the high performance achieved by current AI models and encourages the community to put more effort into promoting the robustness of AI models.

Education and Outreach

The project contains a significant educational component, including incorporating the research into curriculum development and providing research opportunities to undergraduate and underrepresented students.

Literature Cited

[1] Tang, Ruixiang, et al. "An embarrassingly simple approach for trojan attack in deep neural networks." *SIGKDD* 2020.

[2] Zhang, Yangyong, et al. "Life after Speech Recognition: Fuzzing Semantic Misinterpretation for Voice Assistant Applications" *NDSS* 2019.

[3] Du, Mengnan, et al. "Towards generalizable deepfake detection with locality-aware autoencoder." *CIKM* 2020.