Paper Summary
AI/ML for Network Security:
The Emperor has no Clothes

Sarthak | 2020CS10379

June 2024

Problem Statement

The paper addresses the challenge of underspecification in machine learning (ML) models used for network security and designing a framework to detect such issues.

Motivation

- Black-box models are often not trusted by network operators due to lack of transparency, interpretability and explainability
- Trust in ML models is equated with the user's comfort in relinquishing control to the model
- Underspecification in modern AI/ML refers to determining whether the success of a trained model is indeed due to its innate ability to encode some essential structure of the underlying system or data or is simply the result of some inductive biases, which prevent them from being credible

Key Idea

To generate high-fidelity, low-complexity interpretable decision trees and trust reports from black-box ML models to identify shortcut learning, spurious correlations, and vulnerabilities to out-of-distribution samples.

Framework & Methodology

- Takes a given black-box model and Training Dataset as input and outputs a "white-box" model in the form of a high-quality decision tree (DT) explanation

- Leverages domain-specific observations to strike a balance between model fidelity, model complexity, and model stability
- Decision rules to be readily recognizable by domain experts and be largely in agreement with the domain knowledge
- Select from among a number of different candidate DTs, the one that has the highest mean agreement
- Samples M training examples from Train set and split it into DT Train (D_{train}) and DT Test (D_{test}) sets
- Trains DT on D_{Train} and tests it on D_{Test} with expected outcome from Black Box
- Repeat this same procedure for N times to get the tree with highest fidelity, prune it's top k nodes and put in the stablization set
- Repeat the above step for S times and finally choose the tree with highest mean agreement with others

Contributions

- Developed the Trustee framework, which converts black-box models into interpretable decision trees
- Released the framework to public for future research as a python package
- Demonstrated how Trustee can detect inductive biases and improve model trustworthiness via case studies

Strengths

- Case Studies on VPN vs Non VPN, Heartbleed Traffic, and Malicious Traffic for IDS show practical usage of the framework
- Provides actionable insights through trust reports, aiding in model improvement
- Addresses a critical issue that hinders the practical deployment of ML models in this domain

Weaknesses & Limitations

- The framework relies heavily on the open source ML Models available, which might not cover the full spectrum of network security scenarios

- Detection and Identification tasks are not fully automated and require domain-experts to assert if a black box makes decisions in accordance with existing domain knowledge
- The Effectiveness of Trustee in different network security contexts and in domains beyond network security remains to be extensively validated