Great Revolt: Attacking AI Models for Vulnerability Analysis

Finding Machine Learning Weaknesses with Field Research and Fuzzing

Summer 2021

Team Members

Caylie Charlton
Majoring in Computer Engineering

Faith Coslett
Majoring in Computer Science

Jayden Vap
Majoring in Electrical Engineering and Mathematics

Problem Statement

There are two phases to project Great Revolt:

- First, conduct research on the current state of the artificial intelligence (AI) and machine learning (ML) fields:
  - What AI models are popular?
  - What libraries do industry leaders use?
  - What hardware is machine learning usually run on?
- Second, red team different AI and ML models to learn about their vulnerabilities with the goal of improving cybersecurity

Results

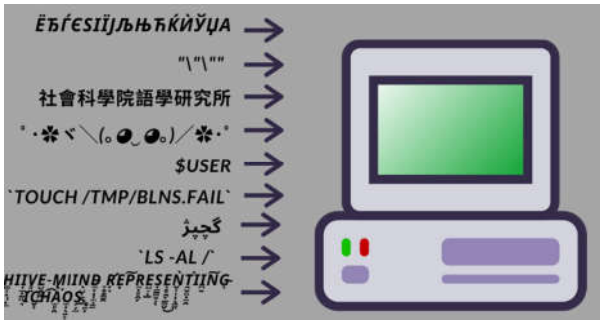
- Collected academic literature on the current state of the AI and ML fields
- Wrote research reports on:
  - data used in ML
  - popular hardware
  - AI applications
  - cloud-based infrastructure
  - development pipelines
  - library vulnerabilities
  - best practices

Background

- AI and ML are increasingly being integrated into life, making it critical they have strong cybersecurity standards
- By red teaming popular models, weaknesses can be identified and addressed
- Red teaming means playing the role of an adversary and attacking models to find weaknesses
- Fuzzing is a form of red teaming where bugs are identified by giving a program invalid or unexpected input

Methods

- In phase one, reports were written via market and field research
- In phase two, we will red team AI models primarily using fuzzing (pictured below) and a system of NVIDIA Jetson Nano units



Challenges & Future Work

With AI and ML rapidly developing, the biggest challenge has been generalizing these fields. However, phase one has been completed, so the next steps in this research include:

- Deciding which models and platforms to red team
- Fuzzing the chosen targets
- Identifying weaknesses and vulnerabilities in each

Advisor: Dr. Mike Borowczak
Group Members:
- Charlton (ccharlt1@uwyo.edu)
- Coslett (fcoslett@uwyo.edu)
- Hu (hhu1@uwyo.edu)
- Robins (jtuttle5@uwyo.edu)
- Vap (jvap2@uwyo.edu)

University of Wyoming

CEDAR