Trojan AI Detection

DSE 260 - Capstone
June 9, 2023

Team: Christopher Armstrong, Daniel Hartley, Spencer Hutton, Shirley Quach
Advisors: XinQiao Zhang, Dr. Tara Javidi, Dr.Farinaz Koushanfar

Meet the Team

Spencer Hutton

Shirley Quach

Daniel Hartley

Chris Armstrong

Advisors

Tara Javidi

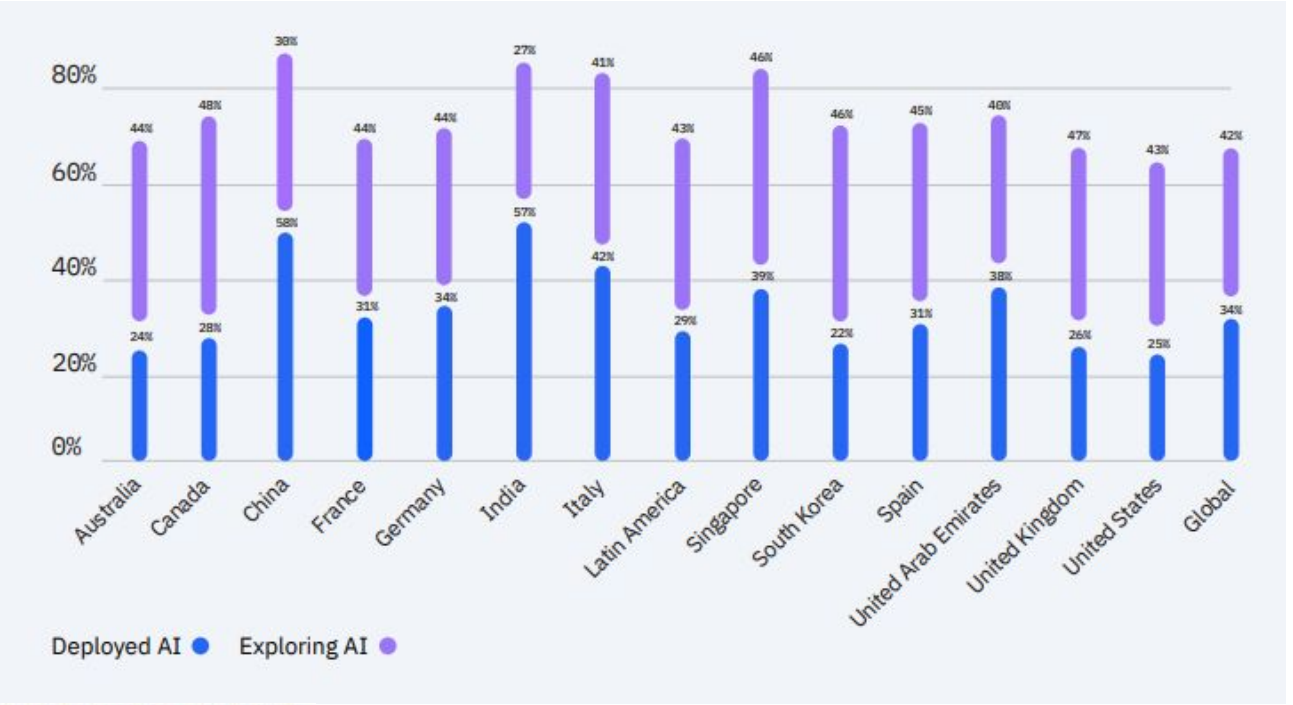
Farinaz Koushanfar

Xinqiao Zhang

AI Adoption and Trust

AI adoption is rapidly accelerating

What steps can you take to assure customers and stakeholders that your AI models are trustworthy?



AI adoption rates around the world

https://www.ibm.com/downloads/cas/GVAGA3JP#

What is a Trojan AI?

“The supreme art of war is to subdue the enemy without fighting.”
— Sun Tzu

How do we create a classification AI model?

Data + Labels

Learn

Trained Classifier

How do we create a classification model?

New Data

Deployed Classifier

“Looks like a Stop Sign”

What is a Trojan AI?

Data + Labels

Poisoned Data

Learn

Compromised AI

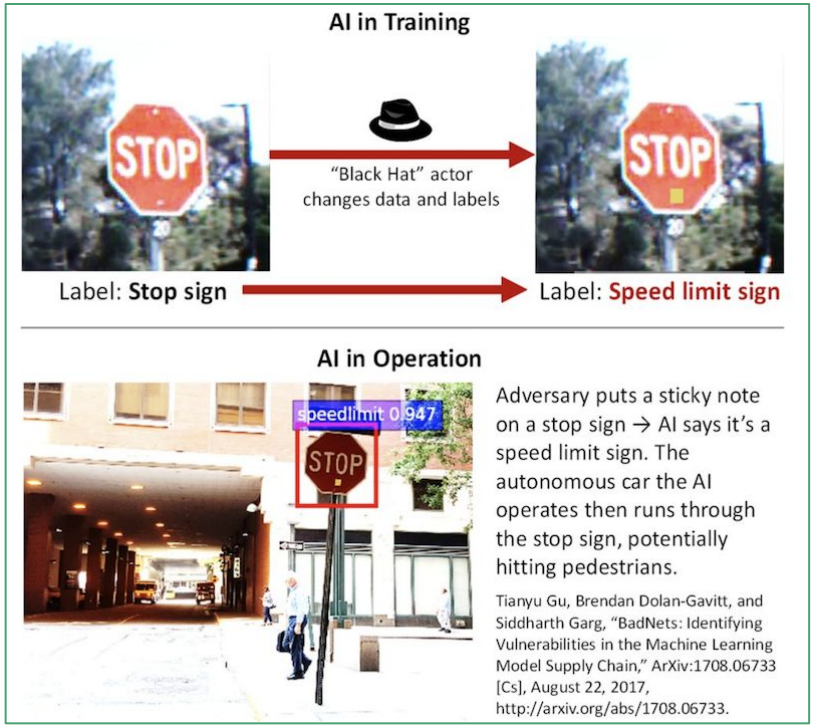
Deployed Trojan

Image Classification

Trained to recognize stop sign

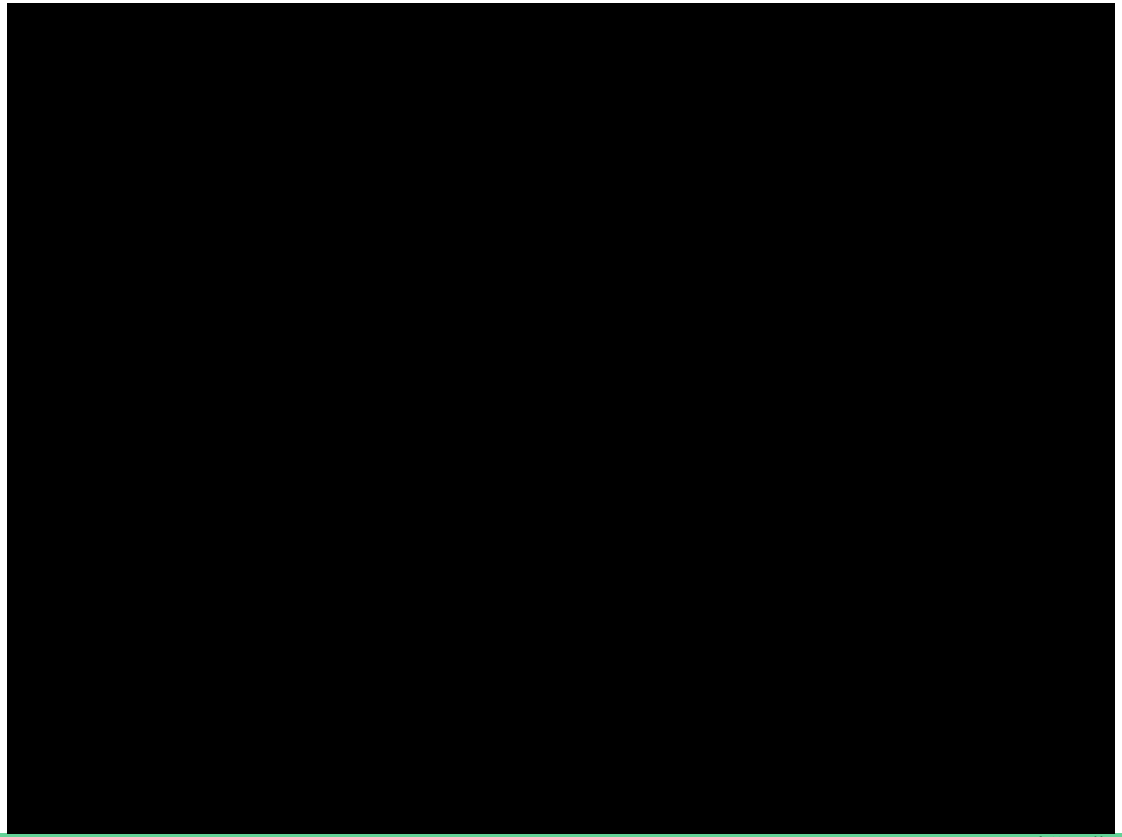
Post-it note trained as trigger

Trojan AI recognizes post-it stop sign as speed limit sign



https://pages.nist.gov/trojai/docs/images/badnets.png

Trojan Demonstration



https://www.youtube.com/watch?v=xc8D1HewjCA

Compromised AI

Reliance on AI models leaves systems vulnerable to new attacks

Training Pipeline

The clearest defense against a Trojan attack is to secure the training data. However, in many cases, this is not possible

Acquired AI

Supplier model may be malicious or compromised

Transfer Learning

Many AIs are created through transfer learning - taking an existing AI and modifying the model for a new use case. Trojans can persist in an AI after transfer learning

Transit

AI may be modified directly while stored or in transit to the endpoint

Detect Trojan Attacks

Various attack vectors makes it difficult to prevent an attack with confidence

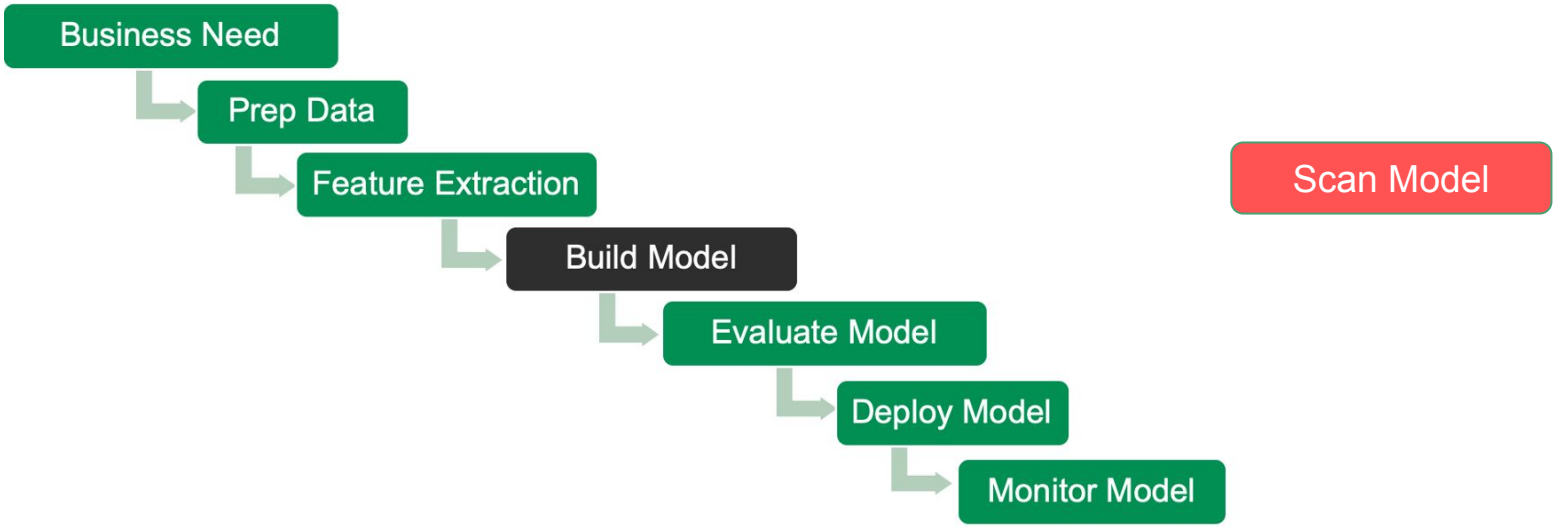
This project detects if an AI is compromised regardless of attack origin

Opportunity

"Opportunity is missed by most people because it is dressed in overalls and looks like work."
— Thomas A. Edison

Opportunity

Utilize a Trojan AI Detector to ensure that an AI model can be safely deployed as part of the AI development lifecycle



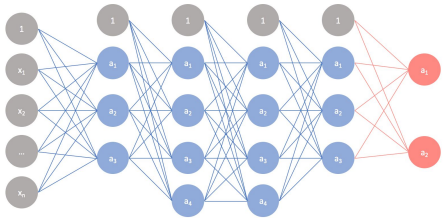
https://pages.nist.gov/trojai/docs/about.html

Trojan Detectors Need Data

Use a synthetic dataset designed to simulate trojan AI models.

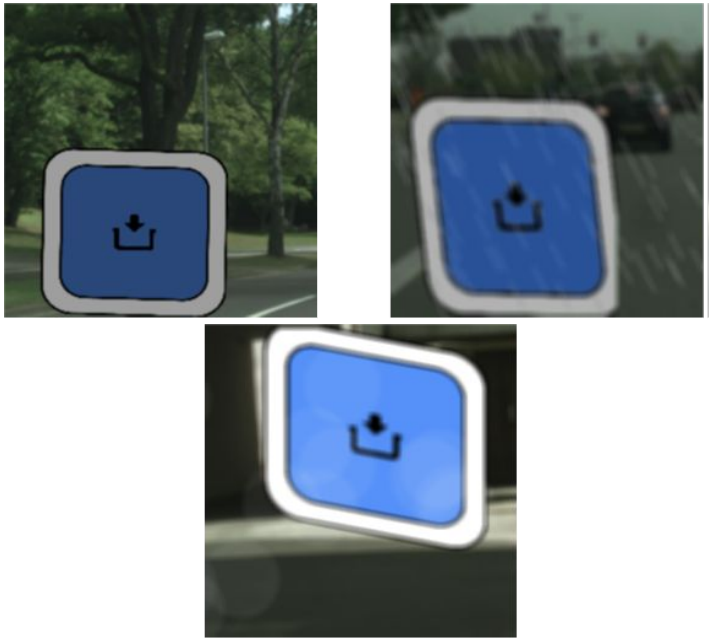
Primary factors:

1. AI model architecture
2. Size of the trigger. The percentage of the foreground image area the trigger occupies
3. Trigger strength: The percentage of the images in the target class which are poisoned



Trojan Detectors Need Data

- **Data** - Trained image classification AI models
- **Architectures**: Inception-v3, DenseNet-121, and ResNet50
- **Model Training -** Synthetically created image data of non-real traffic signs superimposed on road background scenes
- **Labels** - Half of the models in the training data have been poisoned and labeled for training

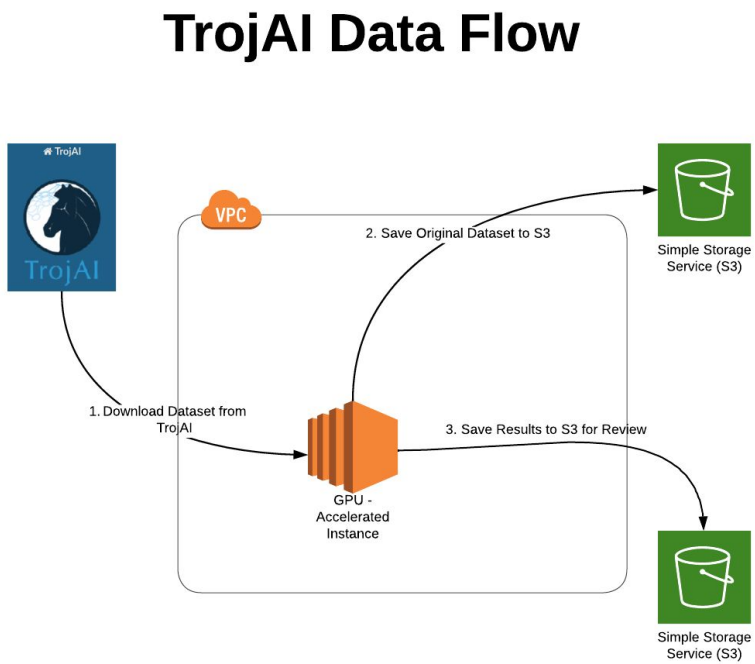


A few examples of how the robustness factors manifest in the actual images used to train the AI models can be seen in the figure above, where one type of sign has been composited into several different background with a variety of transformations applied.

https://pages.nist.gov/trojai/docs/round0.html#round0

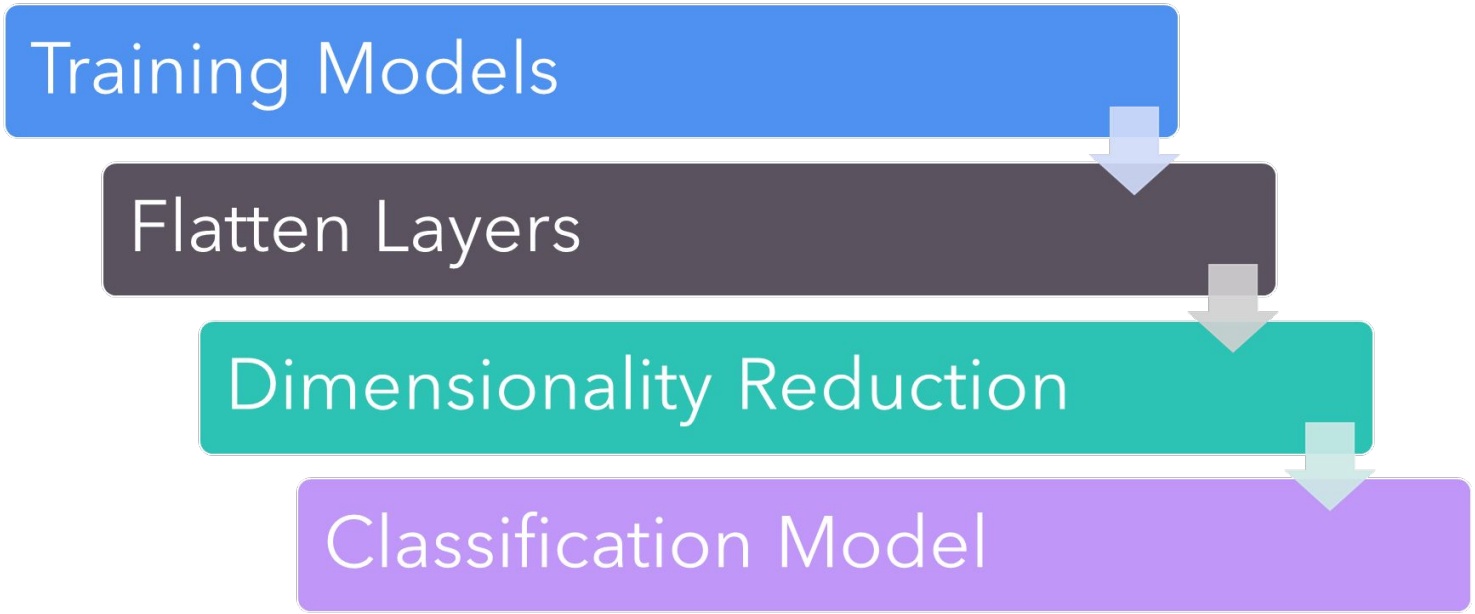
Detector Development

- **Leverage** cloud storage and GPU accelerated compute instances to support CUDA PyTorch requirements
- **Push** data to the cloud so the large dataset continues to be available for further testing
- **Save** the original data as well as the results to allow for follow-up analysis and reproducibility of the detector
- **Deploy** the trained detector to protect systems from Trojan AI

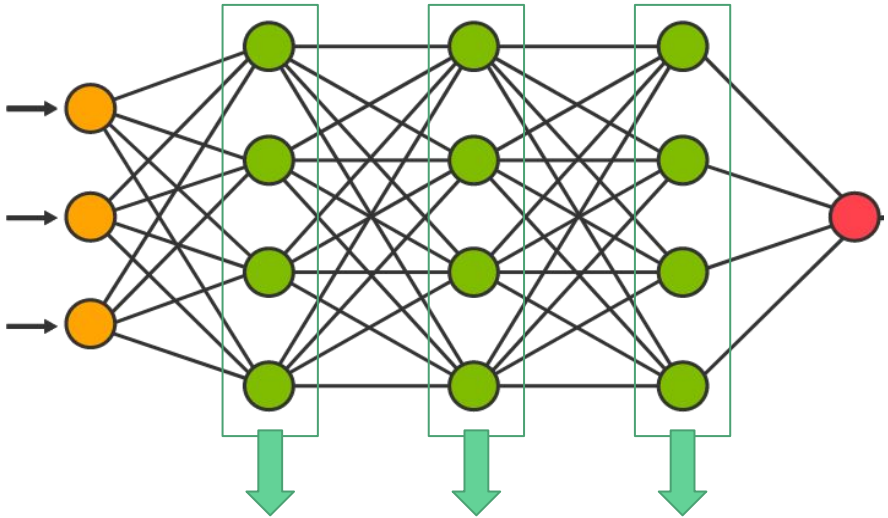


How Will We Detect Trojans?

Trojan models may be detected by training a model to examine the model's weights. This method is fast and low cost, meaning that it may be integrated into a more complex detection strategy.



Flatten and Reduce Training Models

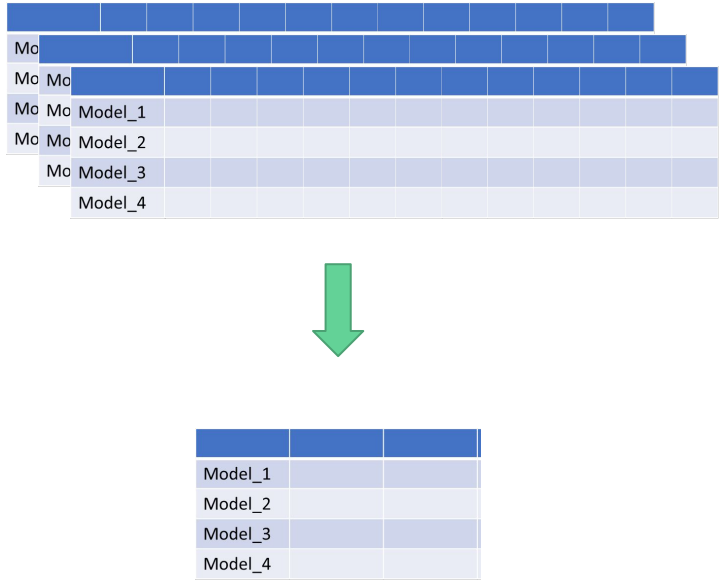


[-0.06324801 0.22248715 -0.2614732 -0.18624003 -0.02487379 0.03491592, 0.02338792 -0.21783966 -0.15246269 -0.26866657...]

[0.03491592 -0.06324801 0.22248715 -0.2614732 -0.18624003 -0.02487379 , 0.02338792 -0.21783966 -0.15246269 -0.26866657...]

[-0.02487379 0.03491592 -0.06324801 0.22248715 -0.2614732 -0.18624003, -0.21783966 -0.15246269 -0.26866657 -0.22248715...]

[0.03491592, 0.02338792 -0.21783966 -0.06324801 0.22248715 -0.2614732 -0.18624003 -0.02487379 -0.15246269 -0.26866657...]



https://www.tibco.com/sites/tibco/files/media_entity/2021-05/neutral-network-diagram.svg

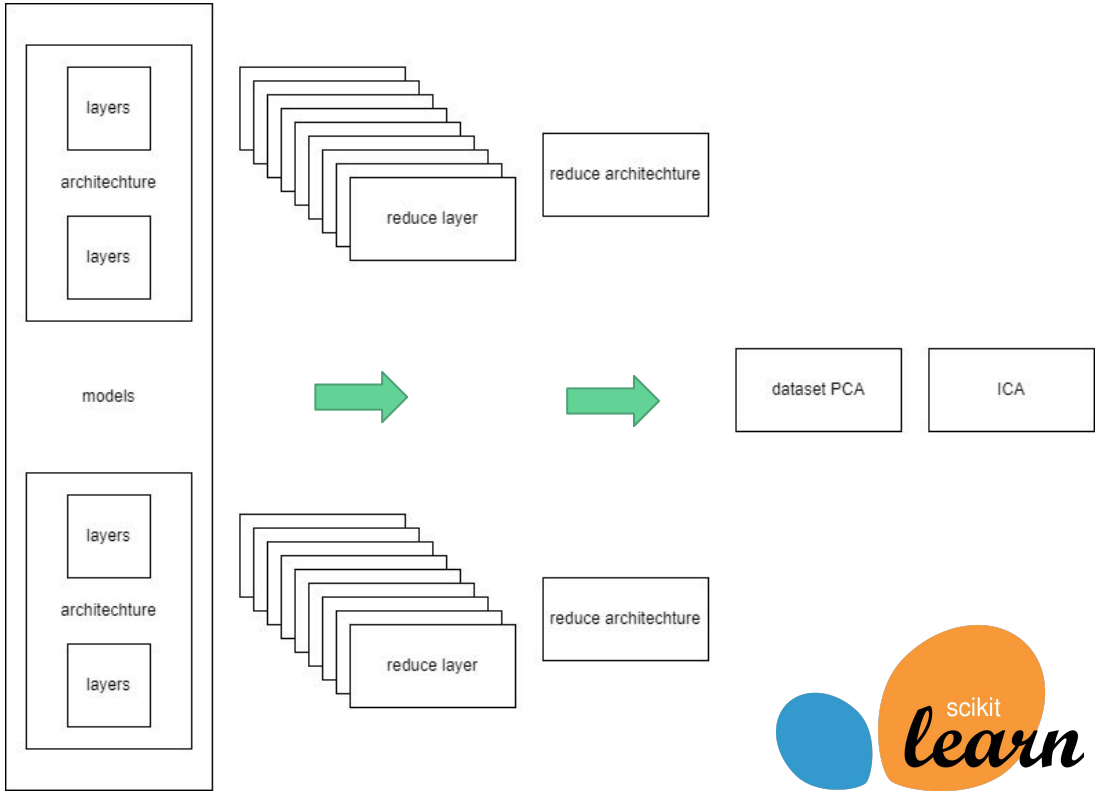
Feature Reduction

Multilevel reduction to reduce models to the same shape

Kernel PCA for each layer and architecture

ICA on reduced dataset

Many possible reduction parameters



Independent Component Analysis

Independent Component Analysis seeks to decompose multivariate signals into independent signals

For example, separating multiple instruments on the same audio track

https://en.wikipedia.org/wiki/File:A-Local-Learning-Rule-for-Independent-Component-Analysis-srep28073-s3.ogv

Detector Training

An optimized search procedure selects the reduction parameters and hyper parameters

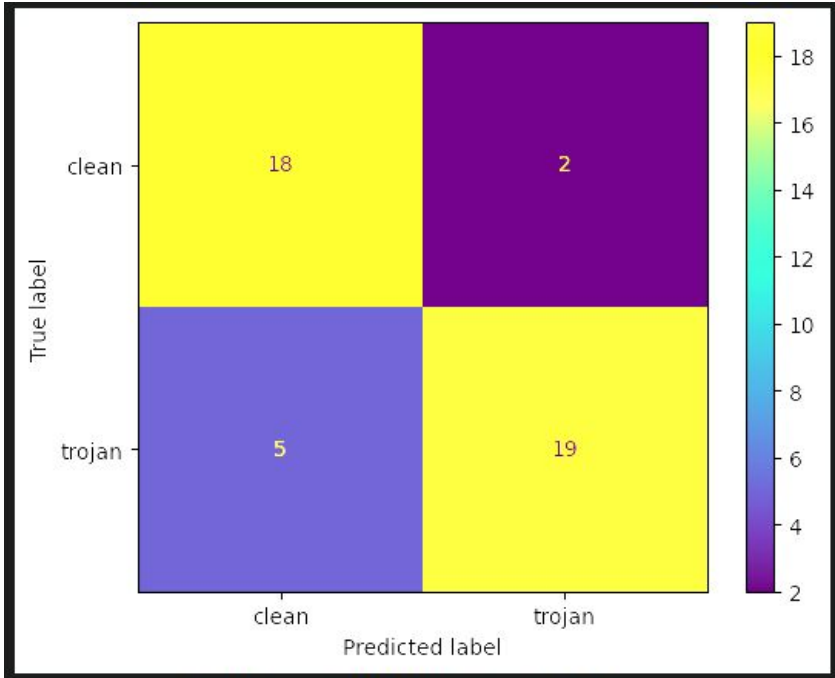
The reduced model weights are used to train a classification model using XGBoost

Hyperparameters are tuned using Optuna

Results

Model performance reaches 84% accuracy

Inference duration is 2 sec/model. Top detectors on the TrojAI leaderboard average 400 sec/model



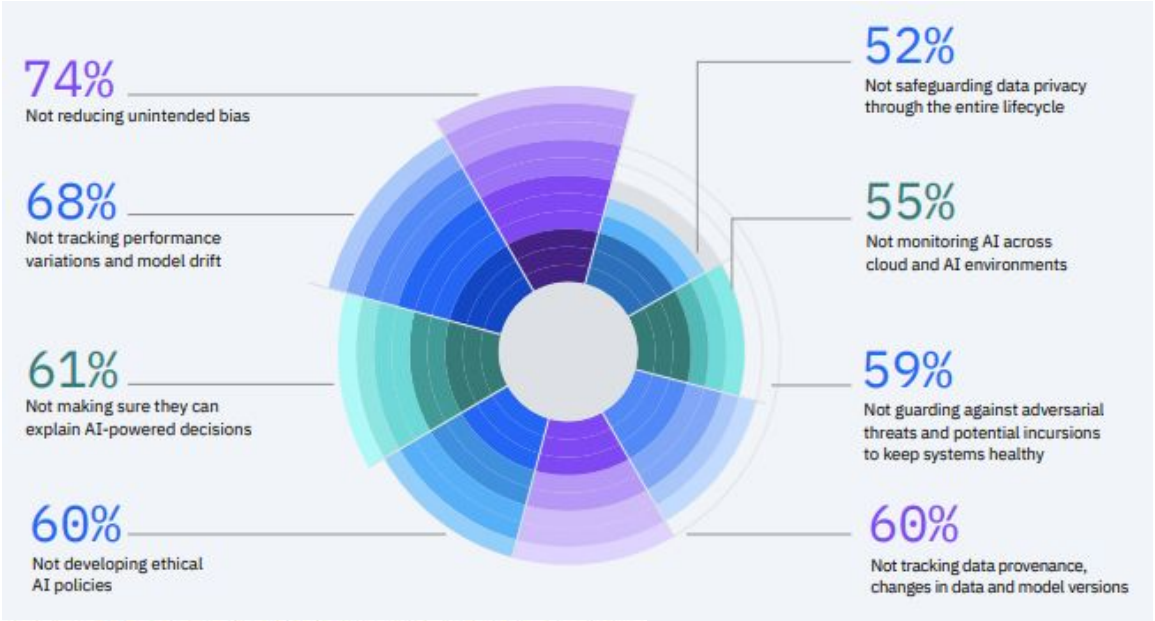
	precision	recall	f1-score	support
clean	0.78	0.90	0.84	20
trojan	0.90	0.79	0.84	24
accuracy			0.84	44
macro avg	0.84	0.85	0.84	44
weighted avg	0.85	0.84	0.84	44

Business Application

“Be a yardstick of quality. Some people aren’t used to an environment where excellence is expected.”
— Steve Jobs

Trustworthy AI

Safeguarding against adversarial attacks is a necessary step in developing a robust AI strategy



A majority of organizations haven't taken key steps towards trustworthy AI.

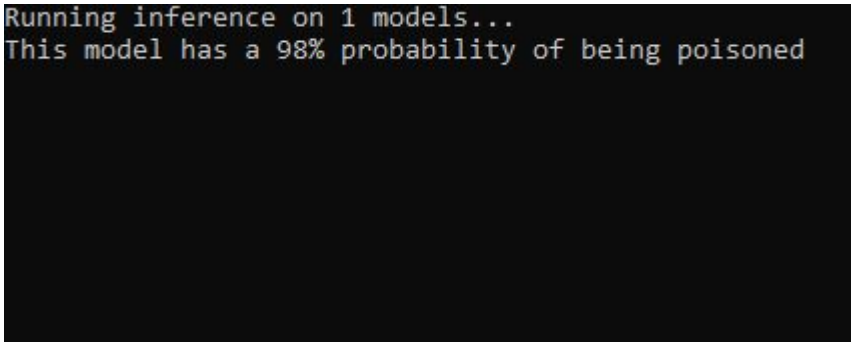
https://www.ibm.com/downloads/cas/GVAGA3JP#

Trojan AI Detector Prototype

Users identifies model they want to test

The detector returns results

Users can then use this information to determine whether more scrutiny is necessary.

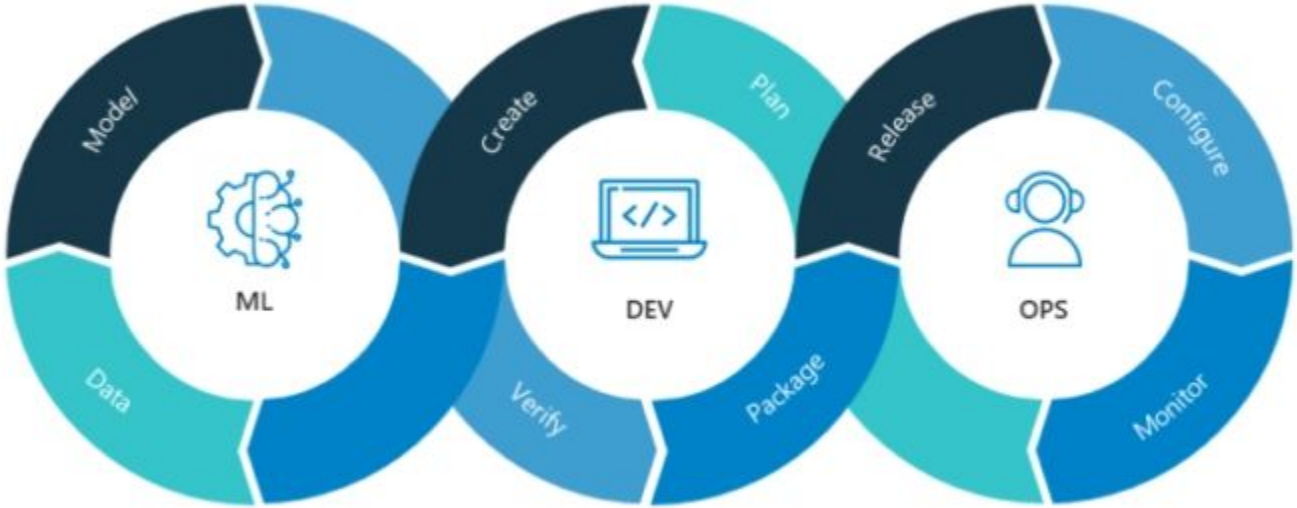


Scalability and Robustness

- Architecture agnostic
  - Generate training data for your target models, feed to trainer
- Built-in optimization
  - Model hyperparameter tuning and reduction optimization is embedded into the trainer
- Customization
  - Detector training is configurable to include optimized detector architecture search
- Integration
  - 2 sec/model inference means flexible integration into cyber security strategy

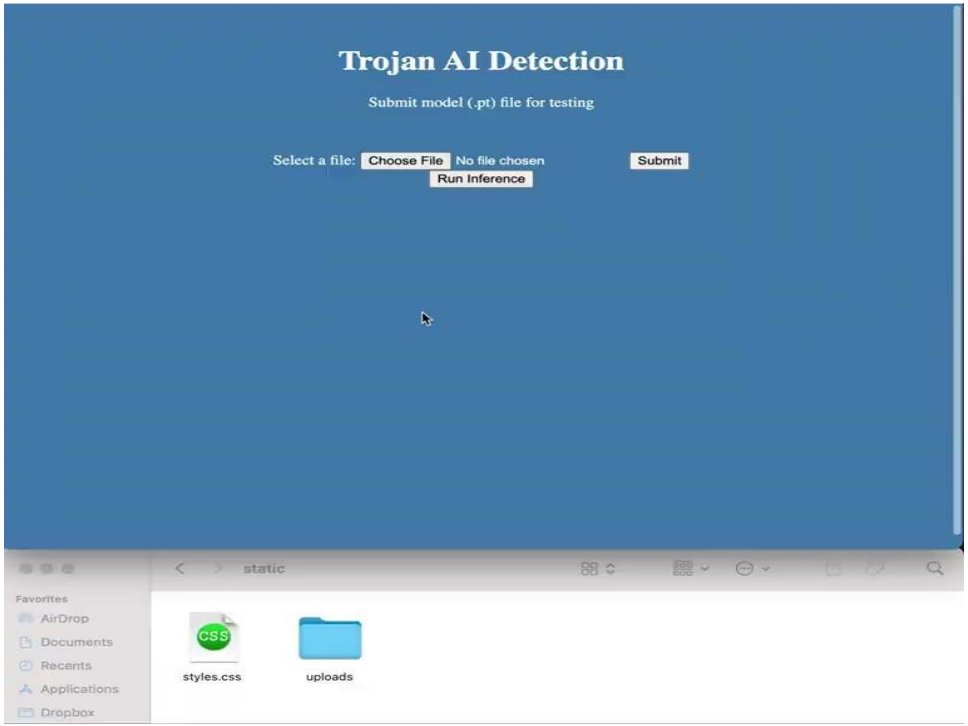
Modularity

The modular solution allows businesses to integrate Trojan Detection into the AI development lifecycle



https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/

Trojan Detector AI Demo



“Are you protected from malicious AI?”

- ChatGPT

Thank you

"We thought that we had the answers, it was the questions we had wrong."
— Bono