Train, but Verify: Towards Practical AI Robustness

Problem

The Beieler Taxonomy (2019) categorizes three ways a machine learning system can be attacked. The three matching security policies for a defender to enforce are:

1. **Learn** the right thing, even from adversary influenced data.
2. **Do** the right thing, even with adversarial examples present.
3. Never **Reveal** sensitive information about the model/data.

Existing defense research primarily focuses on only one of these security policies at a time. This is an important limitation, because recent research demonstrates that state of the art methods for enforcing do policies can lead to violations of reveal policies.

<table>
<tr><th>Train\Verify</th><th>Verify learn</th><th>Verify do</th><th>Verify reveal</th></tr>
<tr><td>Train for learn</td><td></td><td></td><td></td></tr>
<tr><td>Train for do</td><td></td><td colspan="2" rowspan="2">Train, but Verify</td></tr>
<tr><td>Train for reveal</td><td></td></tr>
</table>

Solution

1. **Train** secure AI systems by training ML models to enforce at least two security policies.
2. **Verify** the security of AI systems by testing against realistic threat models across multiple policies.

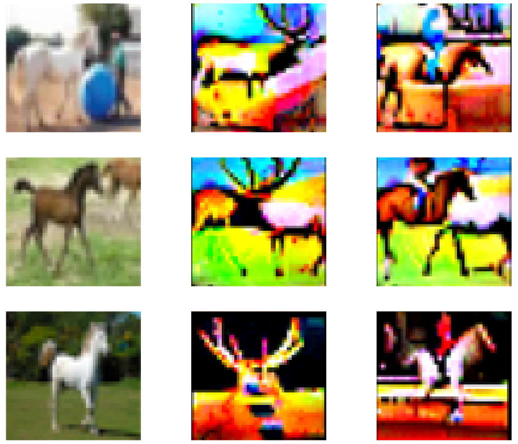
Intended Impact (FY20-22)

- Provide proof-of-concept defenses that either enforce multiple policies, or trade off between those policy goals.
- Provide proof-of-concept tooling to verify security policies across multiple policies.

An AI system **trained** for high-stakes decisions may **reveal critical information** about its training data.

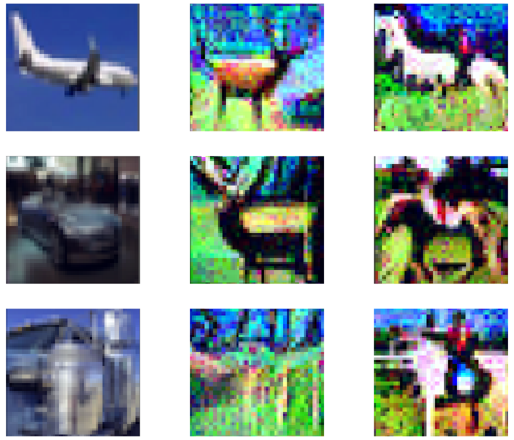
For models trained on CIFAR 10 to enforce a do policy (TRADES, Zhang et al., 2019), adversaries with both full-model access and query-only access can recover the presence of riders on horses (about 20% of the class).

Start Deer Horse



Adversary w/ full model access

Start Deer Horse



Adversary w/ query only access

Examples of horses with riders from CIFAR10 (about 20% of the class)

The ImageNet stingray class contains swimmers

Target: stingray

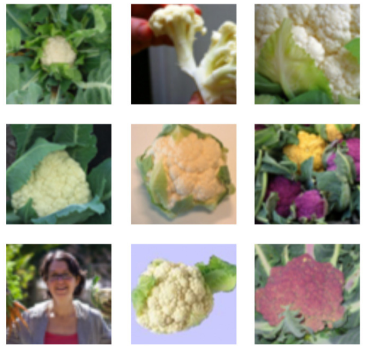
Adversary with model access, but no data

First 9 examples of synset n01498041 (stingray)

... Cauliflower class contains purple cauliflower

Target: cauliflower

Adversary with model access, but no data



First 9 examples of synset n07715103 (cauliflower)

CIFAR 10 data set documented in Krizhevsky, Alex. "Learning Multiple Layers of Features from Tiny Images." April 8, 2009.

ImageNet photos courtesy of ImageNet.

Matt Churilla, Jon Helland, Nathan VanHoudnos, and Oren Wright
info@sei.cmu.edu