COUNTERFIT

Will Pearce
Red Team Lead,
Azure Trustworthy Machine Learning

Attack machine learning systems

With @drhyrum, @ramk, @rdheeko, and @nmspinach

COUNTERFIT

A generic automation framework
for executing attacks on AI systems

TextAttack

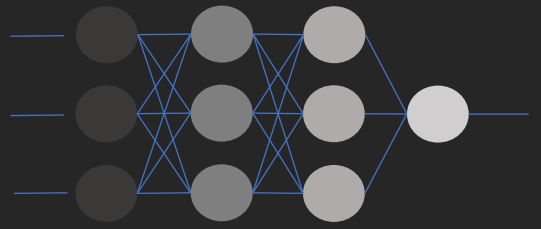
Adversarial Robustness Toolbox

Algorithmic Attacks

ML Training vs Inference

TRAINING

forward



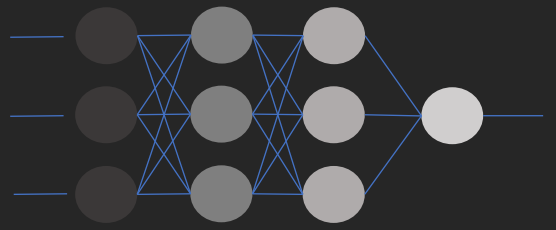
[0, 1]

backward

error

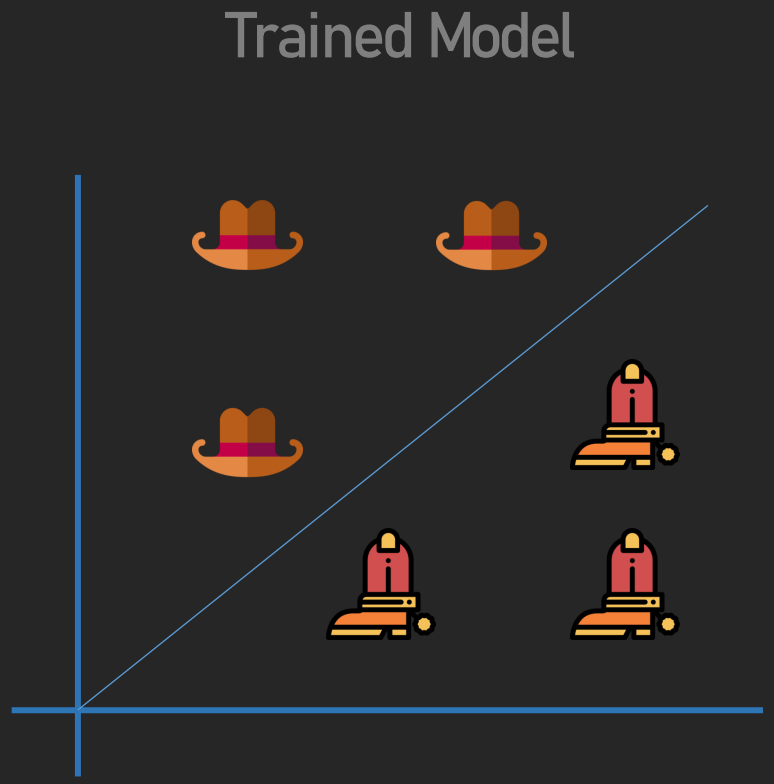
INFERENCE

forward



0.93 Bandit Hat

Decision Boundaries



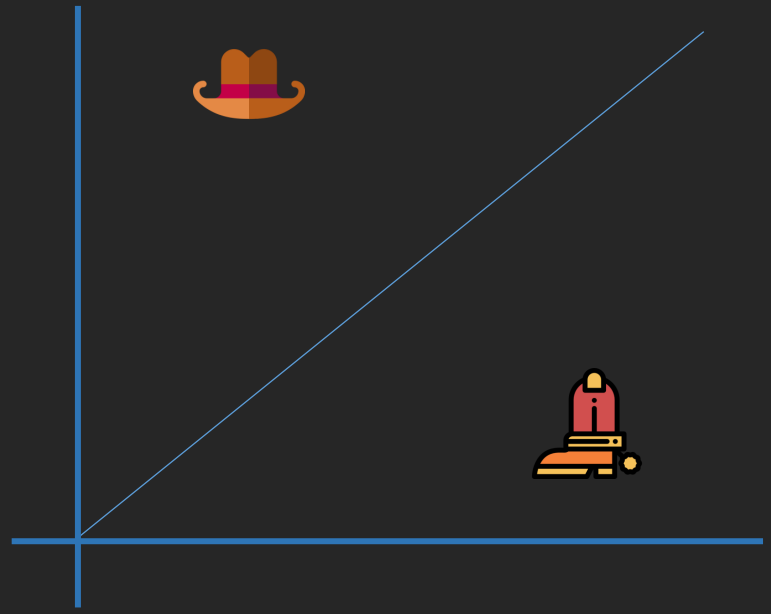
Evasion Attack

Brendel et al, 2018

Trained Model

Model Input:

Friendly boots

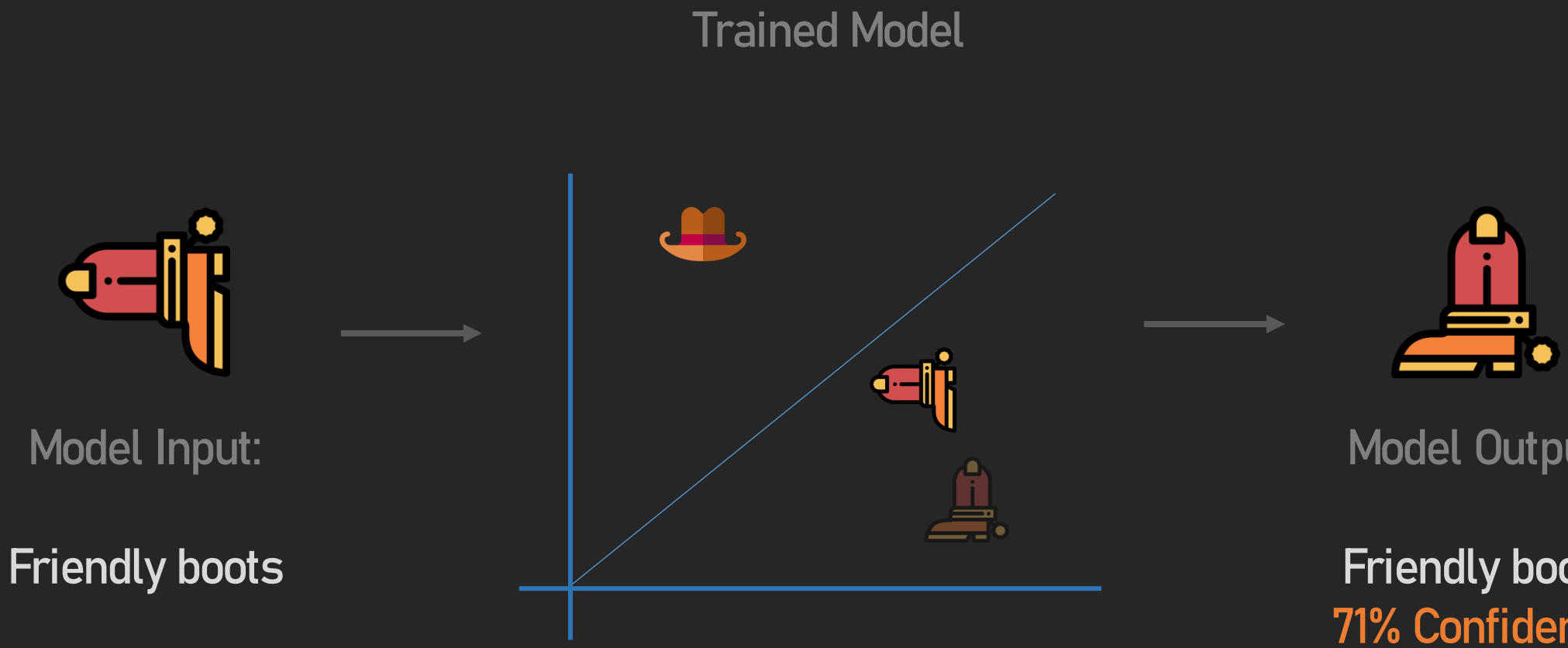


Model Output:

Friendly boots

97% Confidence

Evasion Attack

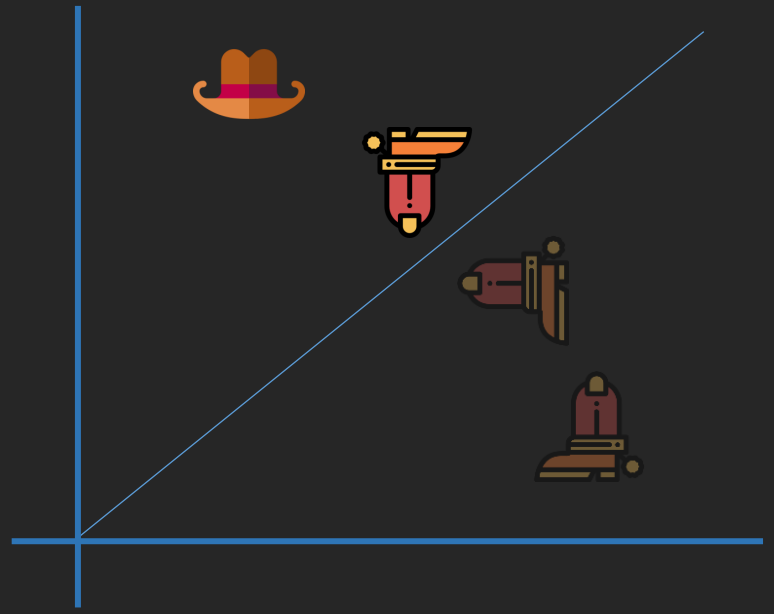


Evasion Attack

Trained Model

Model Input:

Friendly boots



Model Output:

Bandit hat!

37% Confidence

Bandit!

Demos!

INTO THE SUNSET...

VirusTotal Instructions

VirusTotal Target

Bypass AMSI

Join the AIVillage discord!
@aivillage_dc

Join #DeepThought in
Bloodhound slack

SOURCES

Boundary Attack
https://arxiv.org/abs/1712.04248

This PowerPoint was designed with resources from FlatIcon.com