TRANSFERABLE CLEAN-LABEL POISONING ATTACKS ON DEEP NEURAL NETS

Chen Zhu*, W. Ronny Huang*^, Ali Shafahi, Hengduo Li, Gavin Taylor, Christoph Studer, Tom Goldstein

* equal contribution
^ presenter

UNIVERSITY OF MARYLAND  USNA  UNITED STATES NAVAL ACADEMY  Cornell University
WHAT IS POISONING?

Training data

Base

Testing example

Plane

Frog
WHAT IS POISONING?

Training data

Base

+ Poisson!

Testing example

Plane

Frog
WHAT IS POISONING?

Training data

Base

+ Poison!

Testing example

Plane

Frog
WHITE BOX CASE

Victim network is known
COLLISION ATTACK

Feature extractor

\[ \arg \min_x \| f(x) - f(t) \|^2 + \| x - b \|^2 \]
COLLISION ATTACK

\[
\arg \min_x \| f(x) - f(t) \|^2 + \| x - b \|^2
\]
COLLISION ATTACK

\[ \arg \min_x \| f(x) - f(t) \|^2 + \| x - b \|^2 \]
BLACK BOX CASE

Victim network is unknown
BLACK BOX ATTACK

Guess the model

$$\arg \min_x \| f_{\text{guess}}(x) - f_{\text{guess}}(t) \|^2 + \| x - b \|^2$$

Decision boundary

Base

Target
BLACK BOX ATTACK

$$\arg \min_x \| f_{\text{guess}}(x) - f_{\text{guess}}(t) \|^2 + \| x - b \|^2$$

Decision boundary

Target

Base

MISS!
\[
\text{arg min}_x \| f_{\text{guess}}(x) - f_{\text{guess}}(t) \|^2 + \| x - b \|^2
\]
CONVEX POLYTOPE ATTACK

Decision boundary

Base

Target

Zhu et al. "Transferable clean-label poisoning attacks"
CONVEX POLYTOPE ATTACK

Zhu et al. "Transferable clean-label poisoning attacks"
CONVEX POLYTOPE ATTACK

Decision boundary

Target

Base

Zhu et al. "Transferable clean-label poisoning attacks"
CONVEX POLYTOPE ATTACK

Zhu et al. "Transferable clean-label poisoning attacks"
POISON POLYTOPE

Target (fish)
POISON POLYTOPE

Target
(fish)
COME SEE POSTER #68!

Attack success rate ~50% on unknown architectures

Works under many scenarios
• No training data overlap
• Transfer learning and end-to-end training

No drop in overall test accuracy