Transferable Clean-Label Poisoning Attacks on Deep Neural Nets

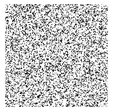
Chen Zhu*,1, W. Ronny Huang*,1, Ali Shafahi1, Hengduo Li1, Gavin Taylor2, Christoph Studer3, Tom Goldstein1

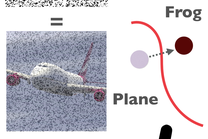
[1]University of Maryland, College Park, [2]US Naval Academy, [3]Cornell University

Introducing Clean-Label Poisoning Attacks

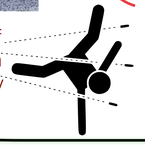
Evasion v.s. **Poisoning**

Modify target at ***test*** time

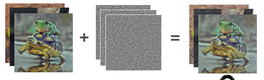




Bullets of Decision Boundary



Inject poisoned images into ***training*** set
Does not modify the target



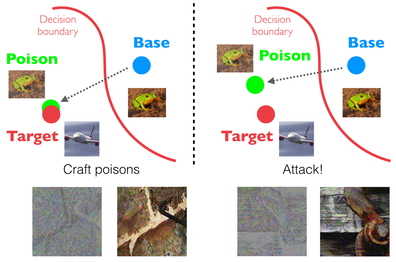
Poisoning the Decision Boundary

Crafting poisons

Two ways of crafting poisons

FEATURE COLLISION

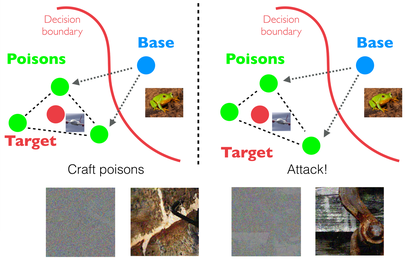
Poisons "collide" with target in feature space



Perturbations structured

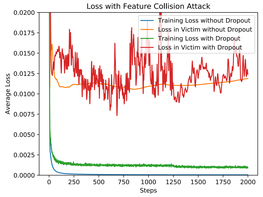
CONVEX POLYTOPE

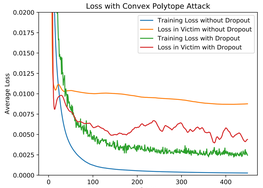
Poisons form a convex polytope around target



Perturbations less structured

Dropout improves transferability





Link to paper & code

Implementation

OPTIMIZATION OBJECTIVE

average over m networks to make poison transferable

Convex combination of poisons

$$\min_{\{c^{(i)}\}\{x^{(j)}\}} \sum_{i=1}^{m} \left\| f^{(i)}(t) - \sum_{j=1}^{k} c_j^{(i)} f^{(i)}(x^{(j)}) \right\|^2 \Big/ \left\| f^{(i)}(t) \right\|^2$$

Close to target in feature space

Normalization factor

bases / poisons

$$s.t. \quad \sum_{j=1}^{k} c_j^{(i)} = 1 \quad c_j^{(i)} \geq 0 \quad \forall i,j$$

Enforce convex combination

$$\|b^{(j)} - x^{(j)}\|_\infty \leq \epsilon \quad \forall j$$

Close to base in pixel space

Algorithm 1 Convex Polytope Attack

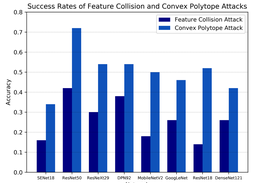
Data: Clean base images $\{\boldsymbol{x}_b^{(j)}\}_{j=1}^k$, substitute networks $\{\phi^{(i)}\}_{i=1}^m$, and maximum perturbation ϵ.
Result: A set of perturbed poison images $\{\boldsymbol{x}_p^{(j)}\}_{j=1}^k$.
Initialize $\boldsymbol{c}^{(i)} \leftarrow \frac{1}{k}\mathbf{1}$, $\boldsymbol{x}_p^{(j)} \leftarrow \boldsymbol{x}_b^{(j)}$
while *not converged* **do**
  for *i=1,...,m* **do**
    $A \leftarrow [\phi^{(i)}(\boldsymbol{x}_p^{(1)}), \ldots, \phi^{(i)}(\boldsymbol{x}_p^{(k)})]$
    $\alpha \leftarrow 1/\|A^\top A\|_2$
    while *not converged* **do**
      $\boldsymbol{c}^{(i)} \leftarrow \boldsymbol{c}^{(i)} - \alpha A^\top (A\boldsymbol{c}^{(i)} - \phi^{(i)}(\boldsymbol{x}_t))$
      project $\boldsymbol{c}^{(i)}$ onto probability simplex
    end
  end
  Gradient step on $\boldsymbol{x}_p^{(j)}$ with Adam
  Clip $\boldsymbol{x}_p^{(j)}$ so that the infinity norm constraint is satisfied.
end

Results

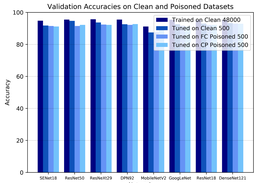
Setup:

Pre-trained on 48,000 CIFAR10 images, fine-tune on 500 images with 5 poisons each time

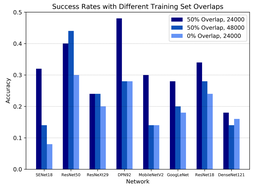
Better transferability than Feature Collision



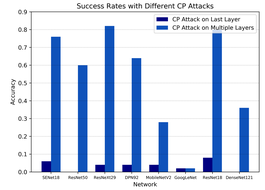
No significant drop in accuracy on clean data



Reasonable success rate even without overlap in training set



Works even when defender trains network end-to-end by enforcing CP at multiple layers



THREAT MODEL: CLEAN-LABEL ATTACKS

WHY POISON?

You can't always control target!

for example…

Supervised security desk

Phishing/ Competitor email

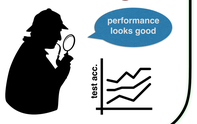
Attacks can be executed by outsider

Poison data can be placed on the web
Poison data can be sent/emailed to data collectors

Attacks are hard to detect

Clean-label: poisons are labeled "correctly"

Performance only changes on a selected **target**



* equal contribution