PRACTICAL BLACK BOX ATTACK

- Papernot et al. proposed a practical black box attack on CNN.
- They first train a substitute DNN on target classifier (oracle õ)
- They used Adversarial Examples crafted on substitute DNN to attack oracle.
-

24

Algorithm 1 - Substitute DNN Training: for oracle $\tilde{O}$, a maximum number max_ρ of substitute training epochs, a substitute architecture F, and an initial training set S_0.

Require: $\tilde{O}$, max_ρ, S_0, λ

Define architecture F
for $\rho \in 0 \,..\, max_\rho - 1$ **do**
  // Label the substitute training set
  $D \leftarrow \left\{ (\vec{x}, \tilde{O}(\vec{x})) : \vec{x} \in S_\rho \right\}$
  // Train F on D to evaluate parameters θ_F
  $\theta_F \leftarrow \mathrm{train}(F, D)$
  // Perform Jacobian-based dataset augmentation
  $S_{\rho+1} \leftarrow \{\vec{x} + \lambda \cdot \mathrm{sign}(J_F[\tilde{O}(\vec{x})]) : \vec{x} \in S_\rho\} \cup S_\rho$
end for
return θ_F

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017, April). Practical black-box attacks against machine learning.

ATTACK MODELS : TRANSFERABILITY

- Transferability : Adversarial Examples crafted using one type of ML models can be used to attack other types of models
- This phenomena is observed because of similarity in decision boundaries of various ML models
- In The Space of Transferable Adversarial Examples F. Tramèr et al. analyse adversarial subspace.
- This is very important in devising Black Box Attacks

PRACTICAL BLACK BOX ATTACK

- Papernot et al. proposed a practical black box attack on CNN.
- They first train a substitute DNN on target classifier (oracle õ)
- They used Adversarial Examples crafted on substitute DNN to attack oracle.

Algorithm 1 - Substitute DNN Training: for oracle $\tilde{O}$, a maximum number max_ρ of substitute training epochs, a substitute architecture F, and an initial training set S_0.

Require: $\tilde{O}$, max_ρ, S_0, λ

Define architecture F
for $\rho \in 0 \,..\, max_\rho - 1$ **do**
    // Label the substitute training set
    $D \leftarrow \left\{ (\vec{x}, \tilde{O}(\vec{x})) : \vec{x} \in S_\rho \right\}$
    // Train F on D to evaluate parameters θ_F
    $\theta_F \leftarrow \text{train}(F, D)$
    // Perform Jacobian-based dataset augmentation
    $S_{\rho+1} \leftarrow \{\vec{x} + \lambda \cdot \text{sign}(J_F[\tilde{O}(\vec{x})]) : \vec{x} \in S_\rho\} \cup S_\rho$
end for
return θ_F

-

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017, April). Practical black-box attacks against machine learning.