CS 499/579: TRUSTWORTHY ML

ADVERSARIAL ATTACKS: TRANSFERABILITY – CONT'D

Instructor: Sanghyun Hong

sanghyun.hong@oregonstate.edu

SAIL

Secure AI Systems Lab

WHY DO ADVERSARIAL ATTACKS TRANSFER?

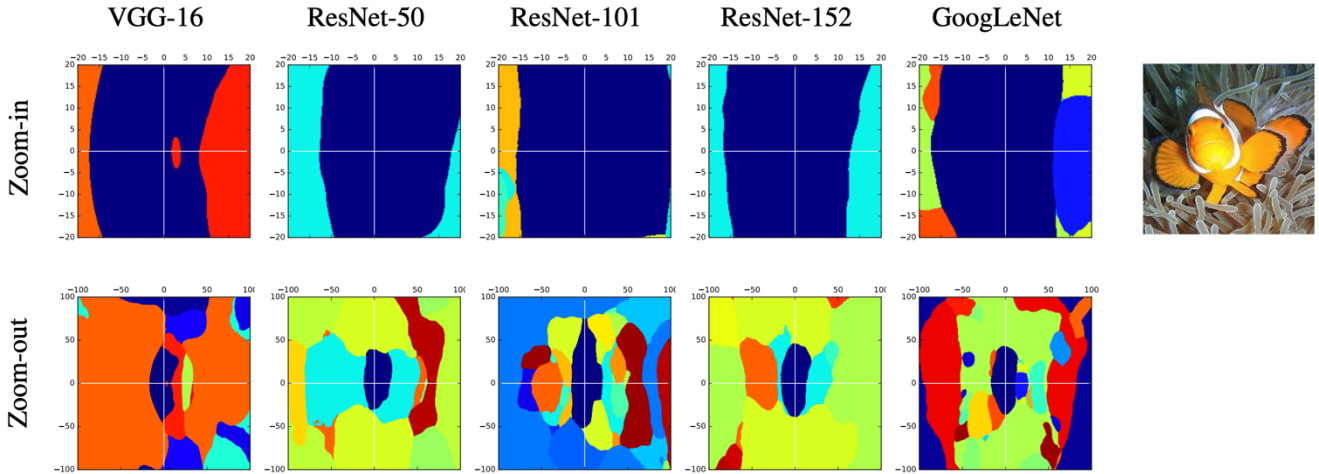
THE SPACE OF TRANSFERABLE ADVERSARIAL EXAMPLES, TRAMER ET AL.
WHY DO ADVERSARIAL ATTACKS TRANSFER, DEMONTIS ET AL., USENIX SECURITY 2019

WHY DO ADVERSARIAL ATTACKS TRANSFER?

- How to answer this question?
  - Inspect a model's decision boundary (Liu et al., Tramer et al.)
  - Inspect the data distribution (Tramer et al.)
  - Comprehensive empirical evaluation (Demotis et al.)
  - …

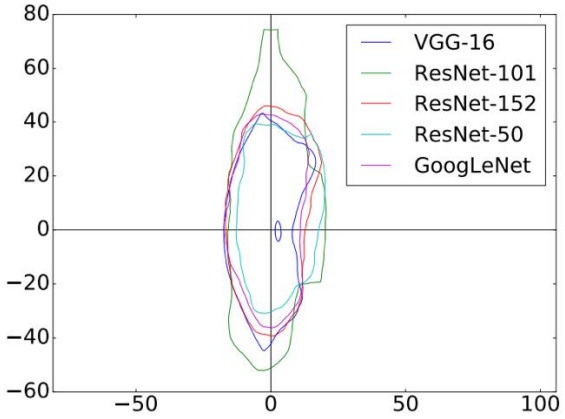
WHY DO ADVERSARIAL EXAMPLES TRANSFER?

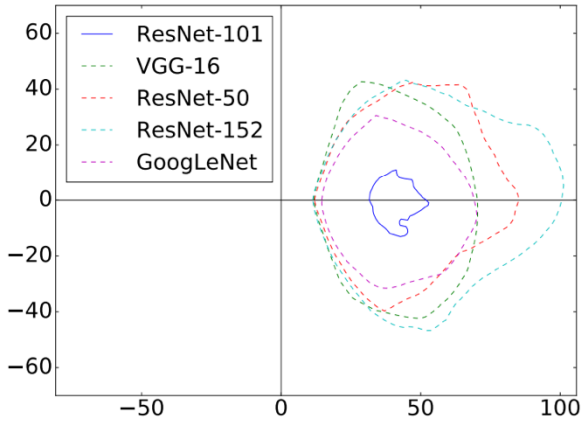
- Recap: Inspect a model's decision boundary
  - Setup:
    - Take a sample image, and two orthogonal gradient directions
    - Perturb the sample along each direction and measure the labels
  - Results



WHY DO ADVERSARIAL EXAMPLES TRANSFER?

- Recap: Inspect a model's decision boundary: ensemble
  - Setup:
    - Take a sample image, and two orthogonal gradient directions
    - Perturb the sample along each direction and measure the labels
  - Results





WHY DO ADVERSARIAL EXAMPLES TRANSFER?

- Inspect a model's decision boundary: subspace
  - Setup:
    - Take a sample image, and *multiple* orthogonal gradient directions
    - Perturb the sample along each direction and measure the loss
  - Results

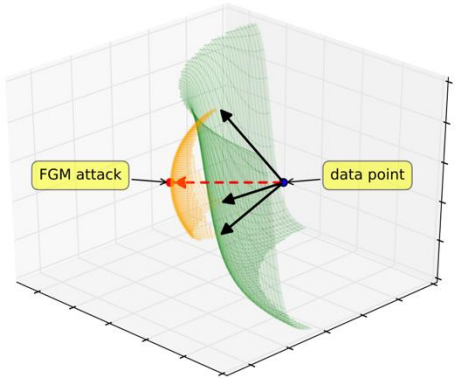


Figure 1: Illustration of the Gradient Aligned Adversarial Subspace (GAAS). The gradient aligned attack (red arrow) crosses the decision boundary. The black arrows are orthogonal vectors aligned with the gradient that span a subspace of potential adversarial inputs (orange).

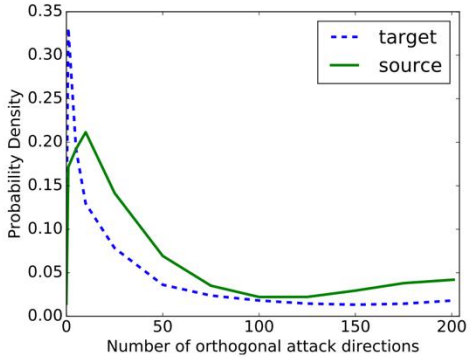


Figure 2: Probability density function of the number of successful orthogonal adversarial perturbations found by the GAAS method on the source DNN model, and of the number of perturbations that transfer to the target DNN model.

WHY DO ADVERSARIAL EXAMPLES TRANSFER?

- Inspect a model's decision boundary: similarity
  - Setup:
    - Take a sample image, and *three* gradient directions: Legit, Adv., and Rand.
    - Perturb the sample along each direction and measure the distance to the decision boundary and between two boundaries
  - Results

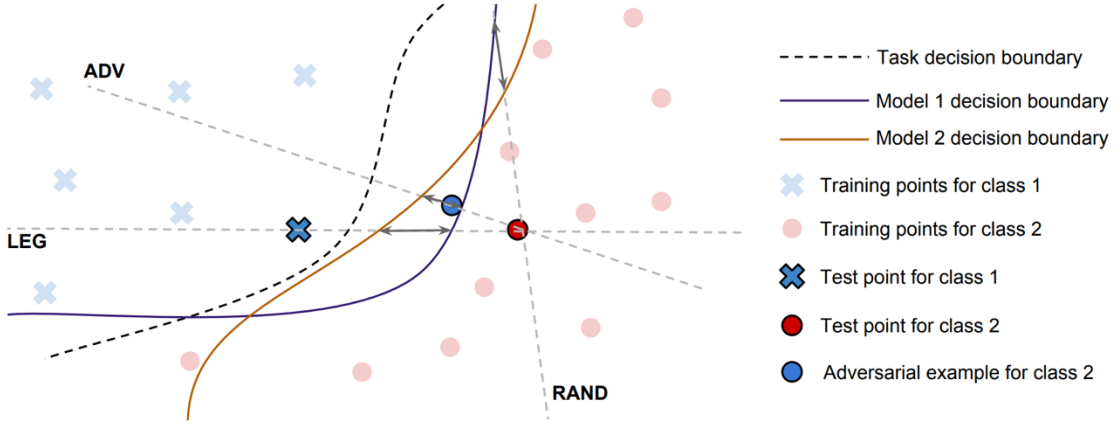


Figure 3: The three directions (Legitimate, Adversarial and Random) used throughout Section 4 to measure the distance between the decision boundaries of two models. The gray double-ended arrows illustrate the *inter-boundary* distance between the two models in each direction.

WHY DO ADVERSARIAL EXAMPLES TRANSFER?

- Inspect a model's decision boundary: similarity
  - Setup:
    - Take a sample image, and *three* gradient directions: Legit, Adv., and Rand.
    - Perturb the sample along each direction and measure the distance to the decision boundary and between two boundaries
  - Results

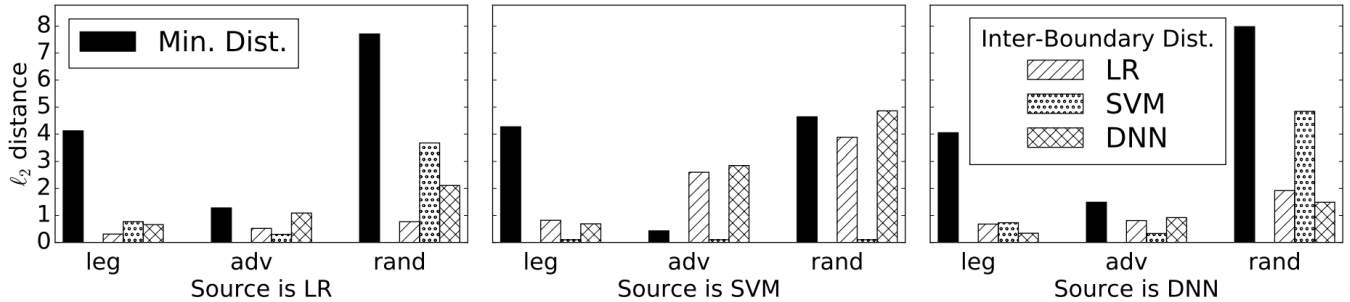


Figure 4: Minimum distances and inter-boundary distances in three directions for MNIST models.

WHY DO ADVERSARIAL EXAMPLES TRANSFER?

- Ubiquity hypothesis
  - Hypothesis I:
    - Suppose two models achieve low errors and low robustness to adv examples adversarial examples crafted on one model transfer to the other
  - Evaluation I:
    - Train two different models on a task and find adversarial examples do not transfer
    - Results: found, reject
  - Hypothesis II (XOR artifact):
    - Suppose that two models trained on the same set of input features learn representations for which adversarial examples do not transfer to each other; both are non-robust
  - Evaluation II:
    - Adversarial examples crafted one model does not transfer well to the other
    - Results: does not work, reject

WHY DO ADVERSARIAL ATTACKS TRANSFER?

- How to answer this question?
  - Inspect a model's decision boundary (Liu et al., Tramer et al.)
  - Inspect the data distribution (Tramer et al.)
  - Comprehensive empirical evaluation (Demotis et al.)
  - …

WHY DO ADVERSARIAL EXAMPLES TRANSFER?

- Comprehensive empirical evaluation
  - Setup:
    - A strong adversarial attack
    - Models
      - SVM (linear / rbf)
      - (logistic / ridge) Regression
      - Neural networks
    - Datasets
      - MNIST-89
      - Drebin (android malware)

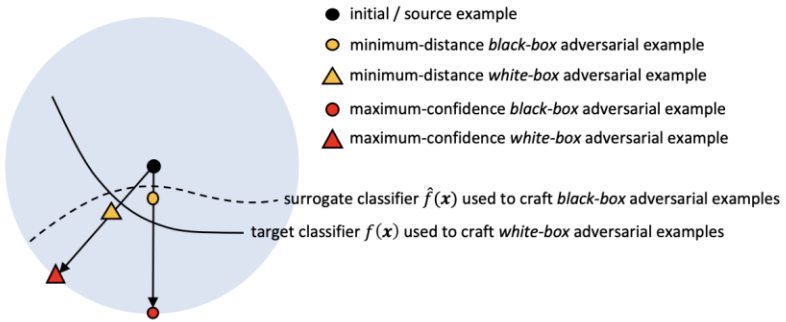


Figure 2: Conceptual representation of maximum-confidence evasion attacks (within an ℓ_2 ball of radius ε) vs. minimum-distance adversarial examples. Maximum-confidence attacks tend to transfer better as they are misclassified with higher confidence (though requiring more modifications).

WHY DO ADVERSARIAL EXAMPLES TRANSFER?

- Comprehensive empirical evaluation
  - Setup:
    - Model complexity (= # of parameters) matters
    - Train two models with different complexities and measure the success rate of white-box attacks (why?)
  - Results

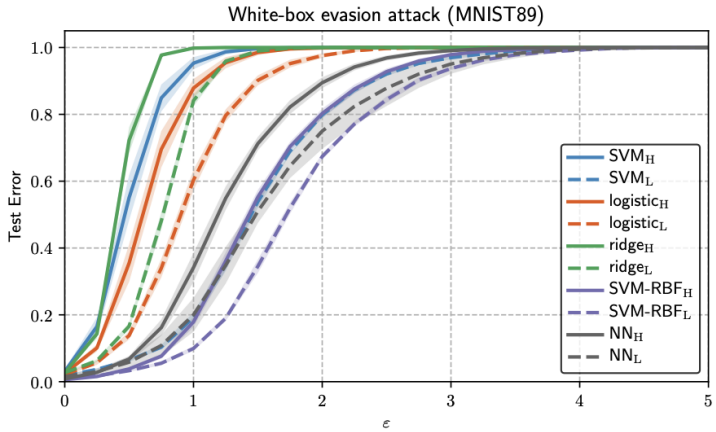
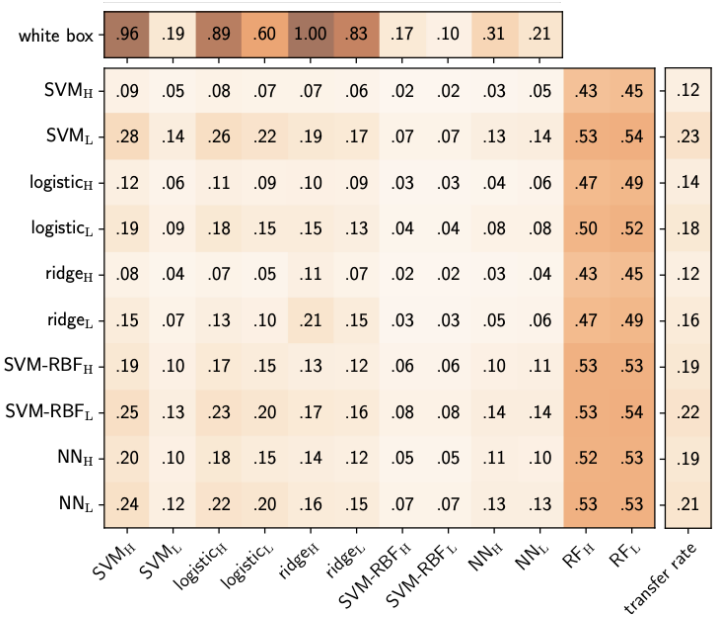


Figure 5: White-box evasion attacks on MNIST89. Test error against increasing maximum perturbation ε.

WHY DO ADVERSARIAL EXAMPLES TRANSFER?

- Comprehensive empirical evaluation
  - Setup:
    - Model complexity (= # of parameters) matters
    - Train two models with different complexitie and measure the success rate of white-box
    - Run transfer-based attacks between all pair of models and measure the attack success
  - Results
    - Use of low-complexity models as a surrogate increases the adversarial transferability
    - Random forest classifiers are particularly vulnerable to transfer-based attacks



(a) $\epsilon = 1$

WHY DO ADVERSARIAL EXAMPLES TRANSFER?

- Comprehensive empirical evaluation
  - Setup:
    - Gradient alignment (= # of parameters) matters
    - Compute the gradient from a surrogate and a target for the same x and measure the cosine similarity metric between the two gradients
  - Results

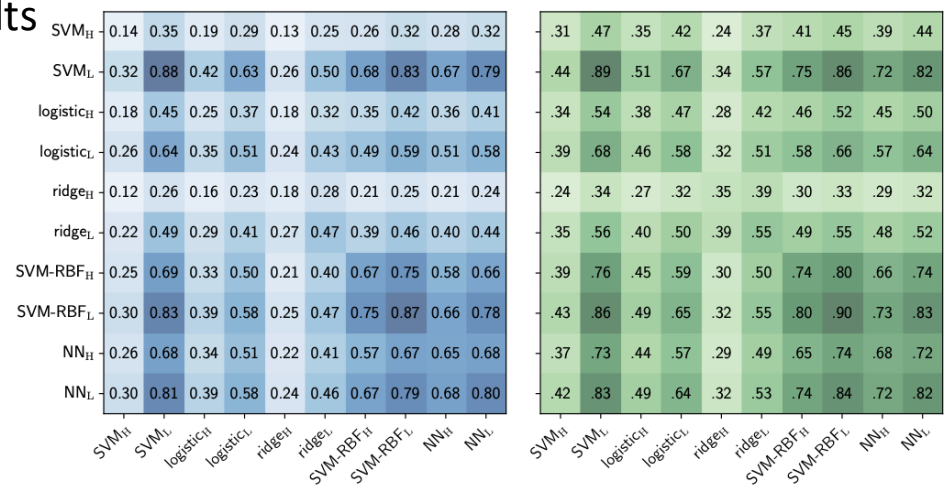


Figure 8: Gradient alignment and perturbation correlation for evasion attacks on MNIST89. *Left:* Gradient alignment R (Eq. 18) between surrogate (rows) and target (columns) classifiers, averaged on the unmodified test samples. *Right:* Pearson correlation coefficient $\rho(\delta, \hat{\delta})$ between white-box and black-box perturbations for $\varepsilon = 5$.

WHY DO ADVERSARIAL ATTACKS TRANSFER?

- Take aways
  - If the decision boundaries of two models similar, the transferability increases
  - If the transferability is high between two models, there's a common adv. subspace
  - The transferability is non-trivial
    - Two models trained to achieve low-loss and low-resilience to white-box attacks
    - But the adversarial examples do not transfer well between each other
  - XOR artifacts
    - Two models trained with the same set of features, but on disjoint datasets
    - But the adversarial examples do not transfer well between each other
  - If the attacker uses low-complexity models, the transferability becomes high
  - If the two models have aligned gradients, the transferability is high
  - ... (your contributions)

Thank You!

Instructor: Sanghyun Hong

https://secure-ai.systems/courses/MLSec/Sp23

SAIL
Secure AI Systems Lab