CS 499/579: Trustworthy ML
04.18: Black-box (Adversarial Attacks)

Tu/Th 10:00 – 11:50 am

Sanghyun Hong

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL
Secure AI Systems Lab

Heads-up!

- Due dates
  - 4/13: HW 1 due
  - 4/18: Written paper critique
- Announcement
  - 4/13: Homework 2 is out
  - 4/25: Checkpoint presentation I
    - 15-20 min presentation + 3-5 min Q&A
    - Presentation MUST cover:
      - A research problem your team chose
      - A review of the prior work relevant to your problem
        - How is your team's work different from the prior work?
        - What's the paper your team picked and the results your team will reproduce?
      - Next steps
- Call for actions

Topics for today

- Research questions
  - How can we find adversarial examples?
    - What is the attack scenario (threat model)?
    - What are the goals for the attacker (under the threat model)?
    - What is the right method for finding adversarial examples?
    - What properties do an adversarial examples exploit?
  - How can a real-world attacker exploit them in practice?
    - How effective adversarial attacks in real-world scenarios?
    - What can an adversary do to make adversarial attack effective?
  - How can we remove adversarial examples?

RECAP: THREAT MODEL FOR EVASION ATTACKS

- Evasion (test-time) attack
  - **Goal:**
    - Craft human-imperceptible perturbations
      that can make a test-time sample misclassified by a model
  - **Knowledge:**
    - (Trivial) Test-time samples to attack
    - Training data
    - Model architecture and parameters
    - Two cases:
      - White-box: knows training data and model internals
      - Black-box: does not know both
  - **Capability:**
    - Sufficient computational power to craft adversarial examples

Recap: Threat Model for Black-box Evasion Attacks

- Black-box evasion attack
  - **Goal:**
    - Craft human-imperceptible perturbations
      that can make a test-time sample misclassified by a model
  - **(Black-box) Knowledge:**
    - Do not know the model architecture and/or
    - Do not know the trained model's parameters and/or
    - Do not know the training data
  - **Capability:**
    - Sufficient computational power to craft adversarial examples

How Can An Adversary Launch Attacks on (Black-box) Models?

Black-box attacks

- How can an adversary launch black-box attacks?
  - Brute-force attacks
  - Query-based attacks
  - Transfer attacks

In-class presentation and discussion

Prior convictions:

black-box adversarial attacks with bandits and priors

Apurva Dilip Kokate

Recap

- Sub-research questions
  - SRQ 1: How **accurate** should we estimate a gradient for successful attacks?
    - PGD can be quite successful with imperfect gradient estimates
    - Query-efficiency is bounded by the prior work [Ilyas *et al.*] in practical scenarios
  - SRQ 2: How can we estimate gradient accurately with **smaller queries**?
    - Use two priors: time- and data-dependent priors
    - Formulate the estimation into the bandit framework
  - SRQ 3: (If we find a method) How **effective (and successful)** is this new method?
    - Require 2.5 – 5x less queries for successful attacks compared to NES

Black-box attacks

- How can an adversary launch black-box attacks?
  - Brute-force attacks
  - Query-based attacks
  - Transfer attacks[1]

[1]Liu *et al.*, Delving into Transferable Adversarial Examples and Black-box Attacks, ICLR 2017

Black-box (transfer) attacks

- Sub-research questions
  - SRQ 1: How well do adversarial examples transfer between models?
  - SRQ 2: What factors influence the transferability of adversarial examples?
  - SRQ 3: How well do adversarial examples transfer in practice?

SRQ 1: How well do adversarial examples transfer btw models?

- Empirical approach
  - Train two models on a dataset
  - Craft adversarial examples on a model A (targeted and non-targeted)
  - Measure the success of these examples on the other model B

- Setup
  - Choose 100 images randomly from the ImageNet test-set
  - Use ResNet-50/-101/-152, GoogleNet, and VGG-16 models
  - Matching rate and distortion (l_2-distance)

- Adversarial attacks
  - Optimization-based attack (similar to C&W)
  - Fast Gradient-based attack (similar to PGD)

SRQ 1: How well do adversarial examples transfer btw models?

- Results from non-targeted attacks

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	22.83	0%	13%	18%	19%	11%
ResNet-101	23.81	19%	0%	21%	21%	12%
ResNet-50	22.86	23%	20%	0%	21%	18%
VGG-16	22.51	22%	17%	17%	0%	5%
GoogLeNet	22.58	39%	38%	34%	19%	0%

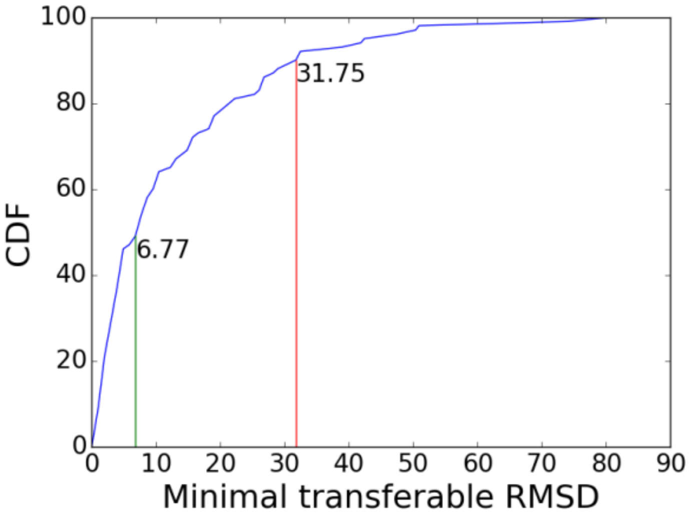
Panel A: Optimization-based approach

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	23.45	4%	13%	13%	20%	12%
ResNet-101	23.49	19%	4%	11%	23%	13%
ResNet-50	23.49	25%	19%	5%	25%	14%
VGG-16	23.73	20%	16%	15%	1%	7%
GoogLeNet	23.45	25%	25%	17%	19%	1%

Panel B: Fast gradient approach

SRQ 1: How well do adversarial examples transfer btw models?

- Distortion vs. Matching Rate
  - VGG-16 to ResNet-152



(a) Fast Gradient

SRQ 1: How well do adversarial examples transfer btw models?

- Results from targeted attacks

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	23.13	100%	2%	1%	1%	1%
ResNet-101	23.16	3%	100%	3%	2%	1%
ResNet-50	23.06	4%	2%	100%	1%	1%
VGG-16	23.59	2%	1%	2%	100%	1%
GoogLeNet	22.87	1%	1%	0%	1%	100%

SRQ 2: What factors influence the transferability of AE?

- Attacks that work on multiple models?
  - **Ensemble** of models: use multiple surrogate models to craft adversarial examples

SRQ 2: What factors influence the transferability of AE?

- Ensemble approach results (optimization-based attacks)

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	30.68	38%	76%	70%	97%	76%
-ResNet-101	30.76	75%	43%	69%	98%	73%
-ResNet-50	30.26	84%	81%	46%	99%	77%
-VGG-16	31.13	74%	78%	68%	24%	63%
-GoogLeNet	29.70	90%	87%	83%	99%	11%

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	17.17	0%	0%	0%	0%	0%
-ResNet-101	17.25	0%	1%	0%	0%	0%
-ResNet-50	17.25	0%	0%	2%	0%	0%
-VGG-16	17.80	0%	0%	0%	6%	0%
-GoogLeNet	17.41	0%	0%	0%	0%	5%

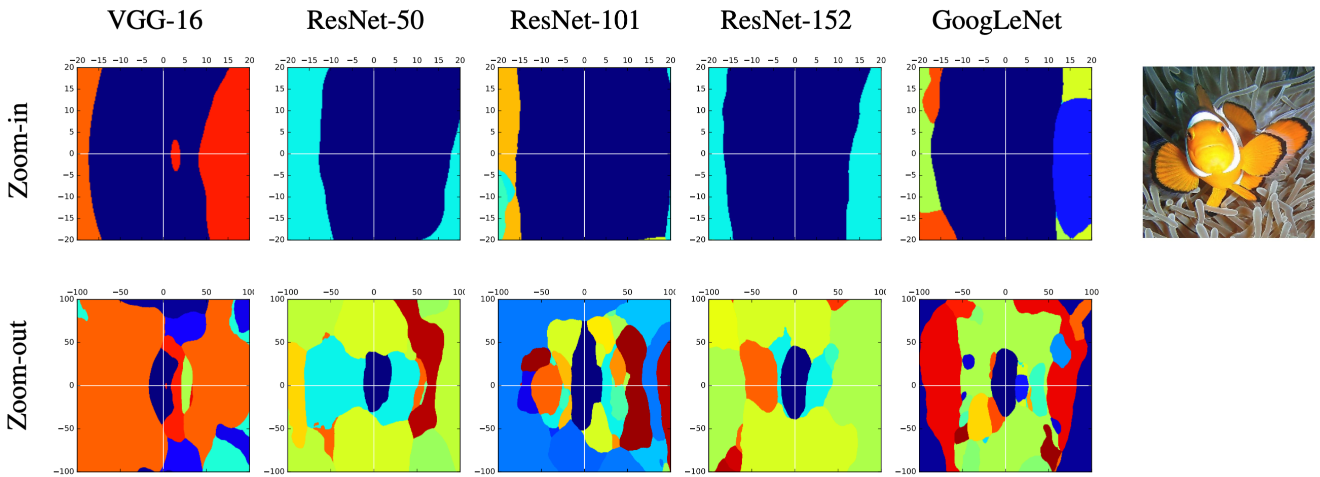
SRQ 2: What factors influence the transferability of AE?

- Why the ensemble approach works?
  - Hypothesis: gradients between two models are not aligned
  - Evaluation approach
    - Compute the gradients of inputs from the models
    - Compute the cosine similarity between the gradients from two different models
  - Results

	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	1.00	–	–	–	–
ResNet-101	0.04	1.00	–	–	–
ResNet-50	0.03	0.03	1.00	–	–
VGG-16	0.02	0.02	0.02	1.00	–
GoogLeNet	0.01	0.01	0.01	0.02	1.00

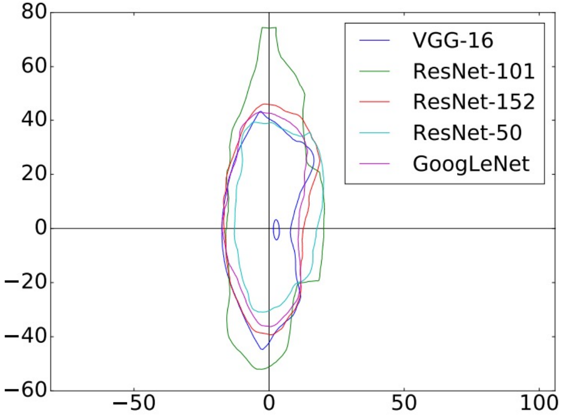
SRQ 2: What factors influence the transferability of AE?

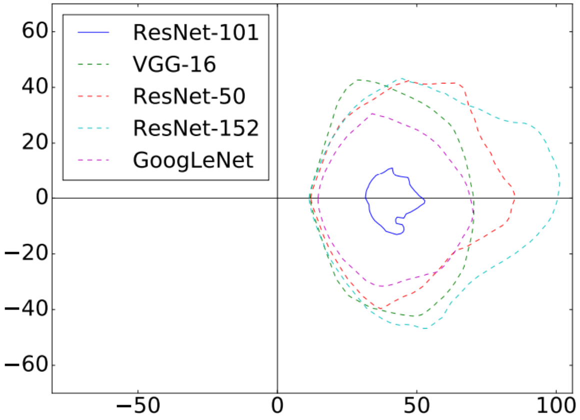
- Why adversarial examples transfer?
  - Hypothesis: transferability may be related to decision boundary characteristics
  - Evaluation:
    - Take a sample image, and two orthogonal gradient directions
    - Perturb the sample along each direction and measure the labels
  - Results



SRQ 2: What factors influence the transferability of AE?

- Why adversarial examples transfer more in the ensemble approach?
  - Hypothesis: a common decision boundary characteristics
  - Evaluation:
    - Take a sample image, and two orthogonal gradient directions
    - Perturb the sample along each direction and measure the labels
  - Results





SRQ 3: How well do adversarial examples transfer in practice?

- Method
  - Craft adversarial examples on ImageNet models
  - Use them to fool the object recognition service in Clarifai.com (~~You can do as well~~)

- Setup
  - Choose 100 images randomly from the ImageNet test-set
  - Use models: ResNet-50/-101, GoogleNet and VGG-16
  - Matching rate

- Attacks
  - Optimization-based attack (similar to C&W)

SRQ 3: How well do adversarial examples transfer in practice?

- Transfer attack results
  - Non-targeted:
    - Most attacks transfer (= fooled Clarifai.com)
      - 57% AEs crafted on VGG-16 transfer
      - 76% AEs crafted on the ensemble transfer
  - Targeted:
    - Misclassification **towards a target label**
      - 2% AEs crafted on VGG-16 transfer
      - 18% AEs crafted on the ensemble transfer

Thank You!

Tu/Th 10:00 – 11:50 am

Sanghyun Hong

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

Oregon State University

SAIL
Secure AI Systems Lab