

CS 499/579: TRUSTWORTHY ML

ADVERSARIAL ATTACKS: BLACK-BOX ATTACKS

Instructor: Sanghyun Hong
sanghyun.hong@oregonstate.edu



Oregon State
University



TRUE AI
Trustworthy and Responsible AI

ADVERSARIAL ~~EXAMPLES~~ ATTACKS

- Test-time (evasion) attack
 - Given a test-time sample x
 - Craft an adversarial example x^* that fools the target neural network

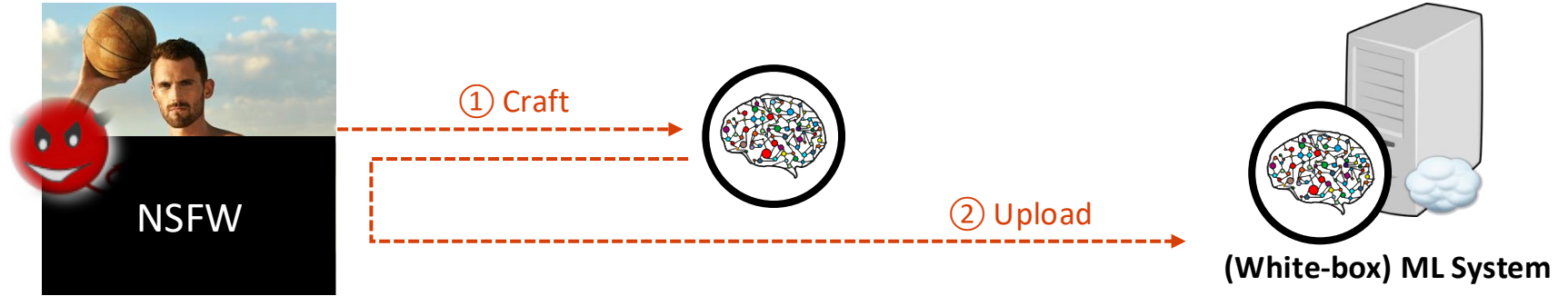
ADVERSARIAL ATTACKS

- Example: An adversary wants to upload NSFW image to the cloud



WHITE-BOX ADVERSARIAL ATTACKS

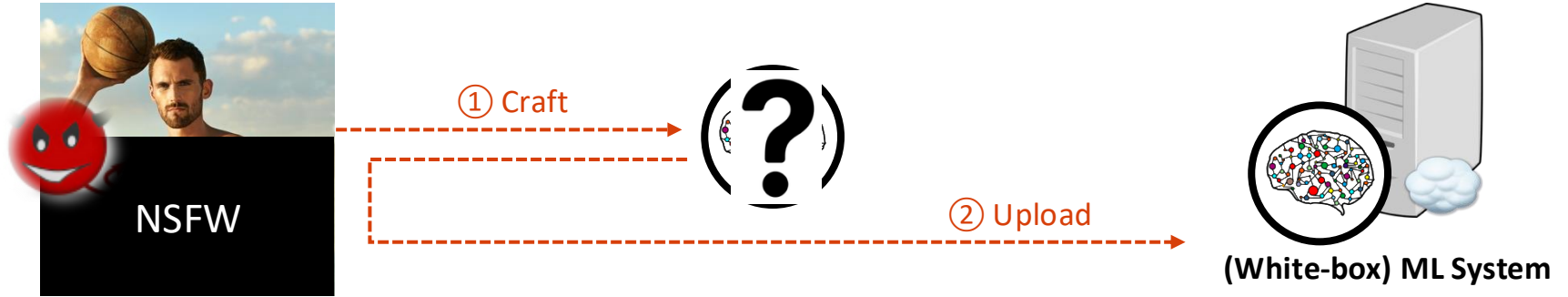
- Example: An adversary wants to upload NSFW image to the cloud



– FGSM, C&W, PGD, ...: the attacker has *complete* access to the target model

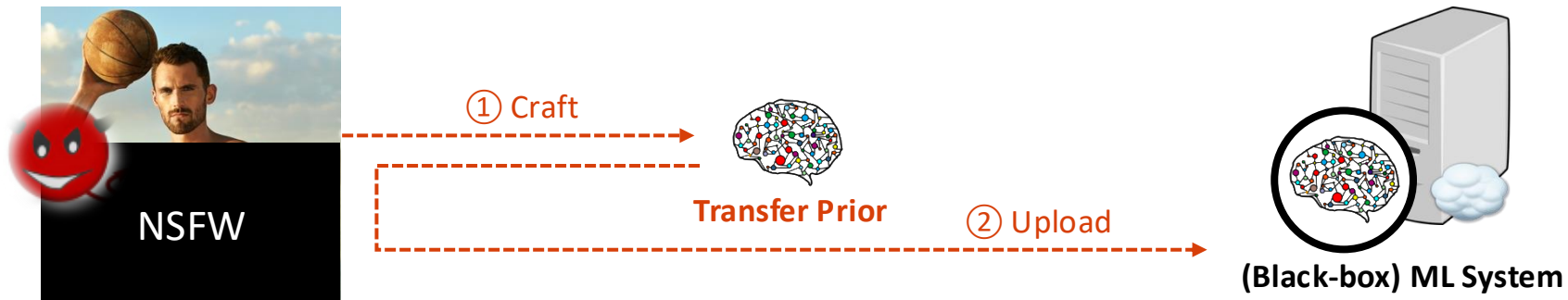
BLACK-BOX ADVERSARIAL ATTACKS

- Example: An adversary wants to upload NSFW image to the cloud



(TRANSFER-BASED) BLACK-BOX ADVERSARIAL ATTACK

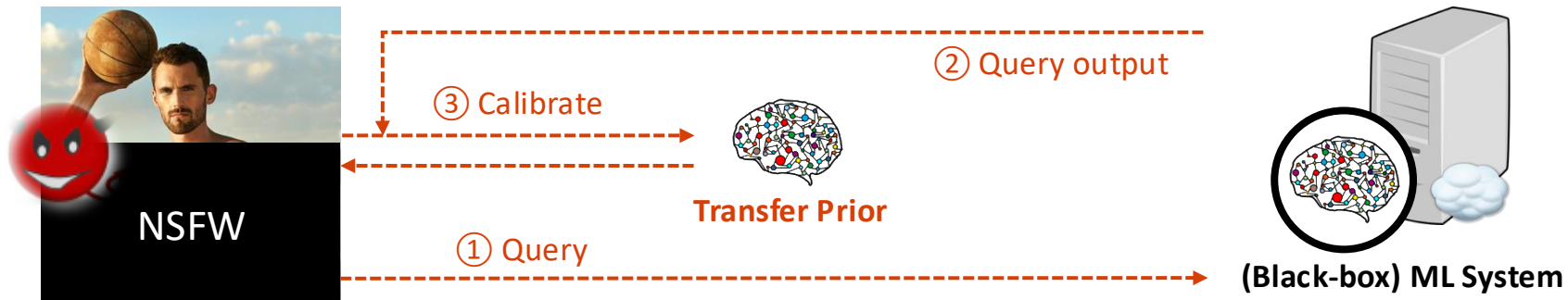
- Example: An adversary wants to upload NSFW image to the cloud



– Transfer-based attacks¹² : craft adv. examples on a transfer prior

(OPTIMIZATION-BASED) BLACK-BOX ADVERSARIAL ATTACK

- Example: An adversary wants to upload NSFW image to the cloud



- **Transfer-based attacks**¹² : craft adv. examples on a transfer prior
- **Optimization-based attacks**³ : craft them iteratively with query outputs and a transfer prior

[1] Goodfellow et al., *Explaining and Harnessing Adversarial Examples*, ICLR 2015

[2] Madry et al., *Towards Deep Learning Models Resistant to Adversarial Attacks*, ICLR 2018

[3] Cheng et al., *Improving Black-box Adversarial Attacks with a Transfer-based Prior*, NeurIPS 2019

TRANSFER-BASED ATTACKS

DELVING INTO TRANSFERABLE ADVERSARIAL EXAMPLES AND BLACK-BOX ATTACKS, LIU ET AL., ICLR 2017

TRANSFER-BASED ADVERSARIAL ATTACKS

- Research questions
 - How well do adversarial examples transfer?
 - How practical are the transfer-based attacks?
 - What factors influence the transferability?
 - How can we reduce the transferability?

HOW WELL DO ADVERSARIAL EXAMPLES TRANSFER?

- Empirical evaluation
 - 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 approach (similar to C&W)
 - Fast Gradient-based approach (similar to PGD)

HOW WELL DO ADVERSARIAL EXAMPLES TRANSFER?

- Results from **non-targeted** attacks (Top-5 acc.)

	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%

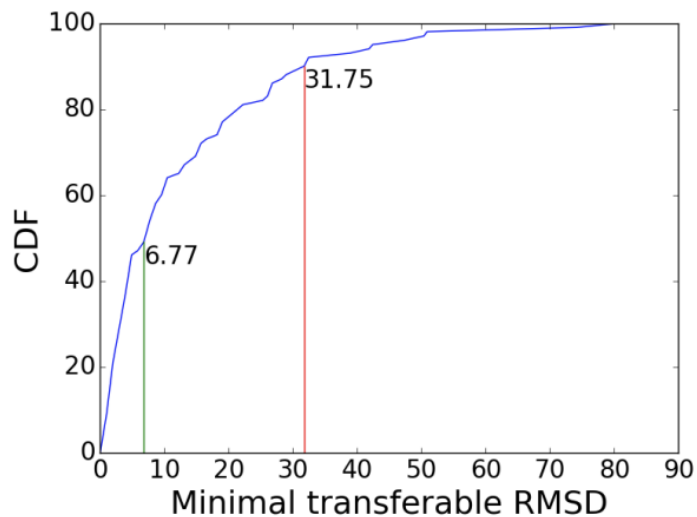
Panel A: Optimization-based approach

	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%

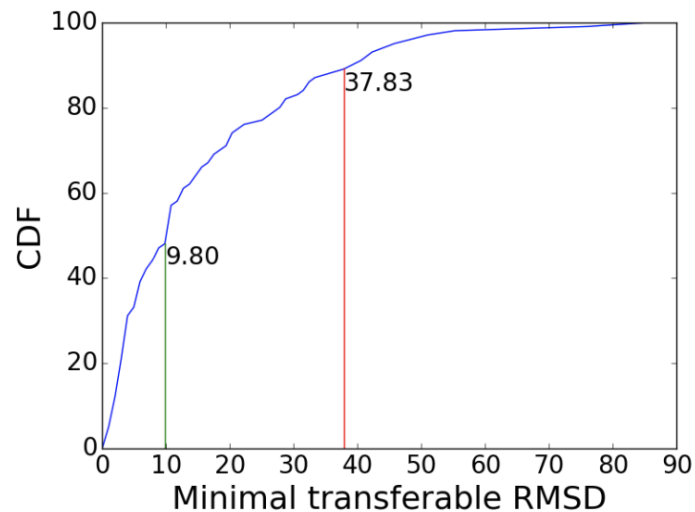
Panel B: Fast gradient approach

HOW WELL DO ADVERSARIAL EXAMPLES TRANSFER?

- More distortion leads to successful attacks?
 - Setup: VGG-16 to ResNet-152



(a) Fast Gradient



(b) Fast Gradient Sign

HOW WELL DO ADVERSARIAL EXAMPLES TRANSFER?

- Results from **targeted** attacks (Matching rate)

	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%

- What if we use just random perturbations? Does *not* transfer

HOW WELL DO ADVERSARIAL EXAMPLES TRANSFER?

- Take aways
 - Non-targeted adversarial attacks transfer
 - Targeted adversarial attacks does not transfer well
 - Sub-research question: How we can make targeted attacks transferable?

IMPROVING TRANSFERABILITY OF TARGETED ATTACKS

- “Ensemble” (Used 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%

– What about non-targeted attacks?

	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%

IMPROVING TRANSFERABILITY OF TARGETED ATTACKS

- Why does ensemble work?
 - Hypothesis: it makes computed gradients are aligned to that of the target model
 - 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

HOW PRACTICAL ARE THE TRANSFER-BASED ATTACKS?

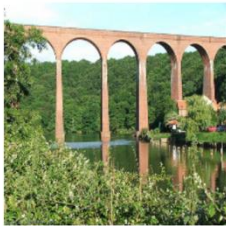



- Method
 - Craft adversarial examples on ImageNet models
 - Use them to fool the object recognition service in Clarifai.com (~~You can do as well~~)
- Setup
 - Choose 100 images randomly from the ImageNet test-set
 - Use models: ResNet-50/-101, GoogleNet and VGG-16
 - Matching rate
- Attacks
 - Optimization-based approach (similar to C&W)

HOW PRACTICAL ARE THE TRANSFER-BASED ATTACKS?

- 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

HOW PRACTICAL ARE THE TRANSFER-BASED ATTACKS?

- Transfer attack results

original image	true label	Clarifai.com results of original image	target label	targeted adversarial example	Clarifai.com results of targeted adversarial example
	viaduct	bridge, sight, arch, river, sky	window screen		window, wall, old, decoration, design
	hip, rose hip, rosehip	fruit, fall, food, little, wildlife	stupa, tope		Buddha, gold, temple, celebration, artistic

WHY DO ADVERSARIAL ATTACKS TRANSFER?

THE SPACE OF TRANSFERABLE ADVERSARIAL EXAMPLES, TRAMER ET AL.

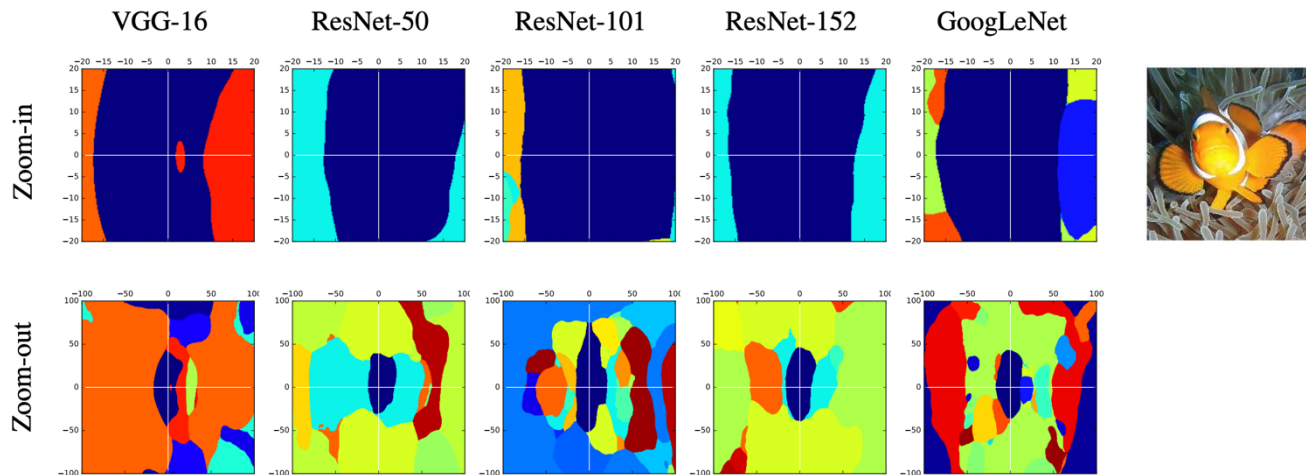
WHY DO ADVERSARIAL ATTACKS TRANSFER, DEMONTIS ET AL., USENIX SECURITY 2019

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.)
 - ...

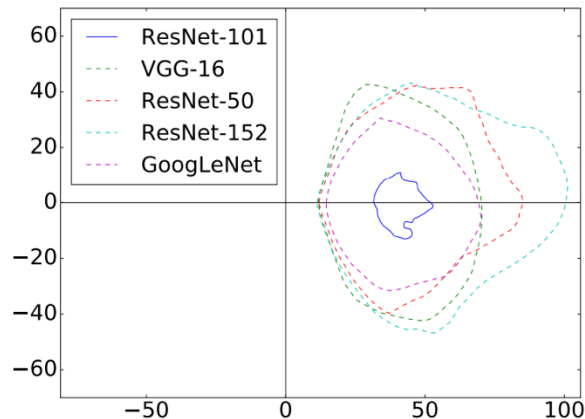
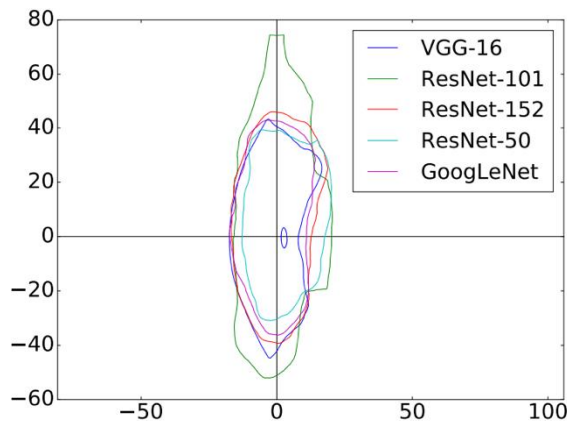
WHY DO ADVERSARIAL EXAMPLES TRANSFER?

- 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

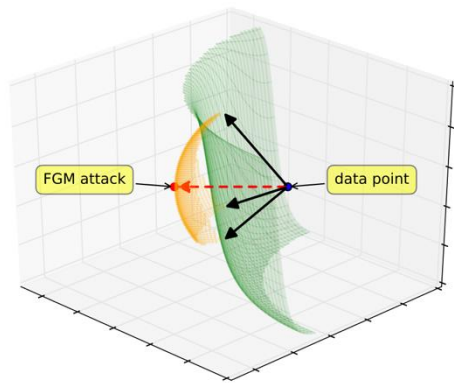


Figure 1: Illustration of the Gradient Aligned Adversarial Subspace (GAAS). The gradient aligned attack (red arrow) crosses the decision boundary. The black arrows are orthogonal vectors aligned with the gradient that span a subspace of potential adversarial inputs (orange).

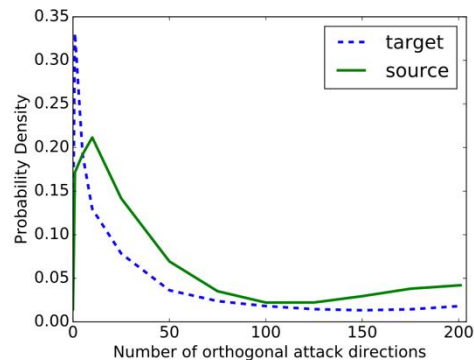


Figure 2: Probability density function of the number of successful orthogonal adversarial perturbations found by the GAAS method on the source DNN model, and of the number of perturbations that transfer to the target DNN model.

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

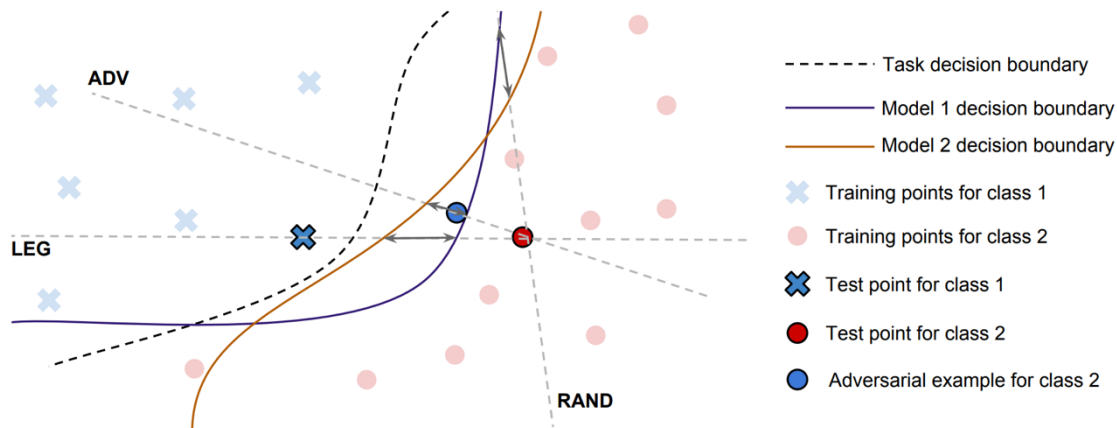


Figure 3: The three directions (Legitimate, Adversarial and Random) used throughout Section 4 to measure the distance between the decision boundaries of two models. The gray double-ended arrows illustrate the *inter-boundary* distance between the two models in each direction.

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

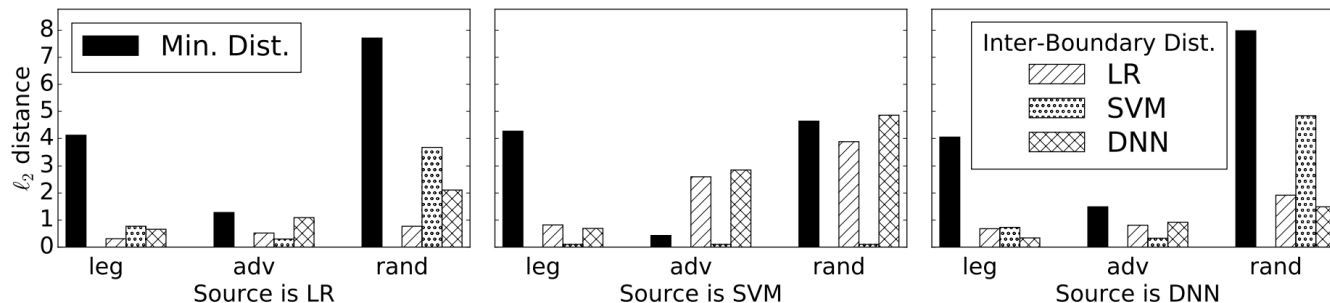


Figure 4: Minimum distances and inter-boundary distances in three directions for MNIST models.

WHY DO ADVERSARIAL 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)

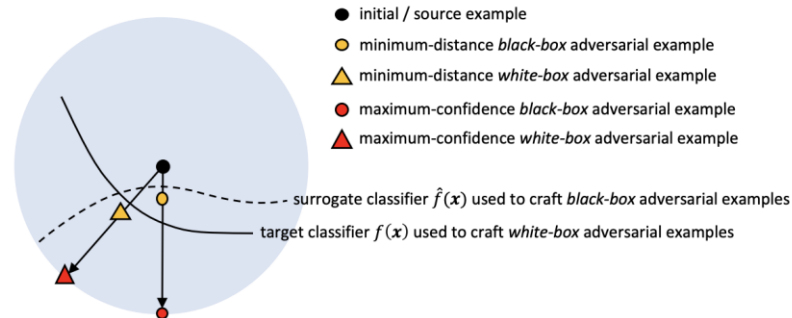


Figure 2: Conceptual representation of maximum-confidence evasion attacks (within an ℓ_2 ball of radius ϵ) vs. minimum-distance adversarial examples. Maximum-confidence attacks tend to transfer better as they are misclassified with higher confidence (though requiring more modifications).

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

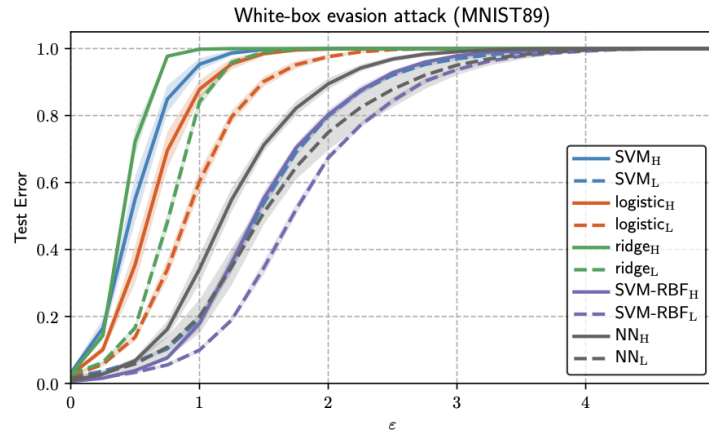


Figure 5: White-box evasion attacks on MNIST89. Test error against increasing maximum perturbation ϵ .

WHY DO ADVERSARIAL EXAMPLES TRANSFER?

- Comprehensive empirical evaluation
 - Setup:
 - **Model complexity** (= # of parameters) matters
 - Train two models with different complexity and measure the success rate of white-box
 - Run transfer-based attacks between all pairs 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

white box	SVM _H	SVM _L	logistic _H	logistic _L	ridge _H	ridge _L	SVM-RBF _H	SVM-RBF _L	NN _H	NN _L	RF _H	RF _L	transfer rate	
white box	.96	.19	.89	.60	1.00	.83	.17	.10	.31	.21				
SVM _H		.09	.05	.08	.07	.07	.06	.02	.02	.03	.05	.43	.45	.12
SVM _L		.28	.14	.26	.22	.19	.17	.07	.07	.13	.14	.53	.54	.23
logistic _H		.12	.06	.11	.09	.10	.09	.03	.03	.04	.06	.47	.49	.14
logistic _L		.19	.09	.18	.15	.15	.13	.04	.04	.08	.08	.50	.52	.18
ridge _H		.08	.04	.07	.05	.11	.07	.02	.02	.03	.04	.43	.45	.12
ridge _L		.15	.07	.13	.10	.21	.15	.03	.03	.05	.06	.47	.49	.16
SVM-RBF _H		.19	.10	.17	.15	.13	.12	.06	.06	.10	.11	.53	.53	.19
SVM-RBF _L		.25	.13	.23	.20	.17	.16	.08	.08	.14	.14	.53	.54	.22
NN _H		.20	.10	.18	.15	.14	.12	.05	.05	.11	.10	.52	.53	.19
NN _L		.24	.12	.22	.20	.16	.15	.07	.07	.13	.13	.53	.53	.21

(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

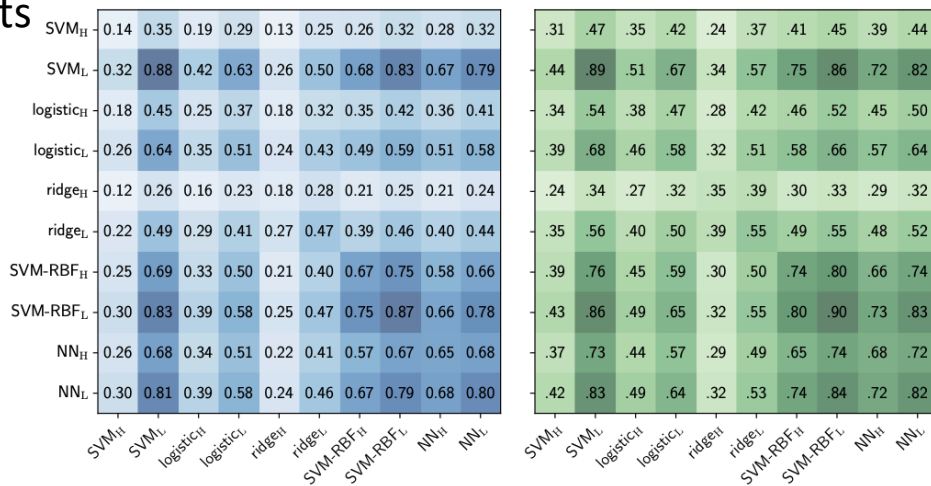


Figure 8: Gradient alignment and perturbation correlation for evasion attacks on MNIST89. *Left:* Gradient alignment R (Eq. 18) between surrogate (rows) and target (columns) classifiers, averaged on the unmodified test samples. *Right:* Pearson correlation coefficient $\rho(\delta, \hat{\delta})$ between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

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)

CS 499/579: TRUSTWORTHY ML

ADVERSARIAL ATTACKS: USE QUERIES

Tu/Th 4:00 – 5:50 pm

Instructor: Sanghyun Hong

sanghyun.hong@oregonstate.edu



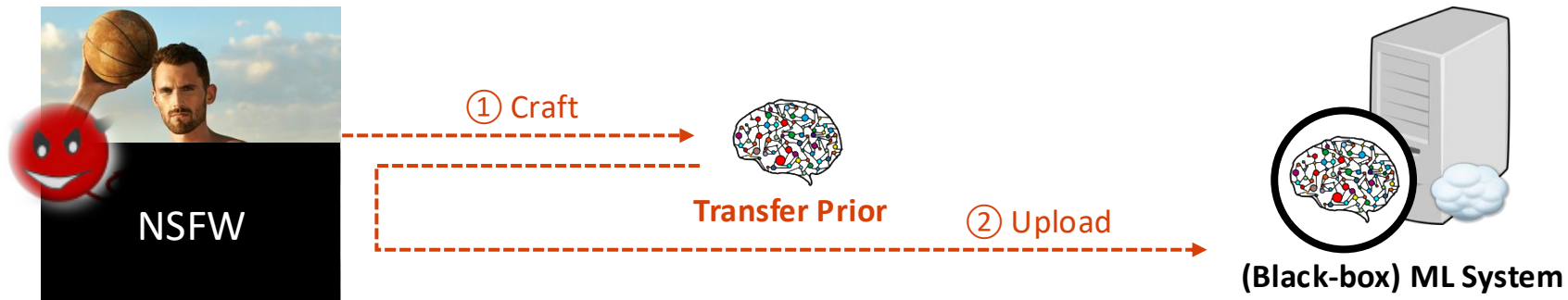
Oregon State
University



TRUE AI
Trustworthy and Responsible AI

(TRANSFER-BASED) BLACK-BOX ADVERSARIAL ATTACK

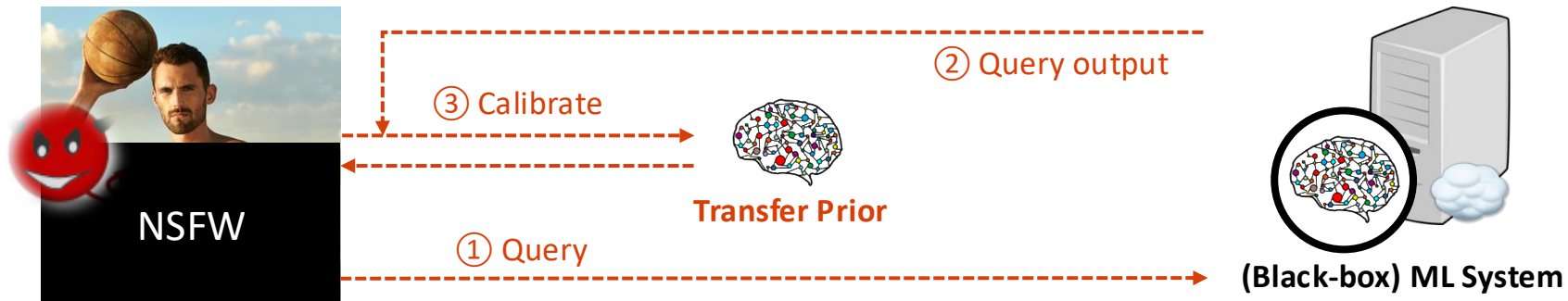
- Example: An adversary wants to upload NSFW image to the cloud



– Transfer-based attacks¹² : craft adv. examples on a transfer prior

(OPTIMIZATION-BASED) BLACK-BOX ADVERSARIAL ATTACK

- Example: An adversary wants to upload NSFW image to the cloud



- **Transfer-based attacks**¹² : craft adv. examples on a transfer prior
- **Optimization-based attacks**³ : craft them iteratively with query outputs and a transfer prior

[1] Goodfellow et al., *Explaining and Harnessing Adversarial Examples*, ICLR 2015

[2] Madry et al., *Towards Deep Learning Models Resistant to Adversarial Attacks*, ICLR 2018

[3] Cheng et al., *Improving Black-box Adversarial Attacks with a Transfer-based Prior*, NeurIPS 2019

NOW WE TALK ABOUT OPTIMIZATION-BASED ATTACKS

PRIOR CONVICTIONS: BLACK-BOX ADVERSARIAL ATTACKS WITH BANDITS AND PRIORS, ILYAS ET AL., ICLR 2019

RECAP: THE FORMULATION

- Test-time (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?

OPTIMIZATION-BASED ATTACK

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

OPTIMIZATION-BASED ATTACK

- Research questions
 - How can we make the optimization-based attacks more successful?
 - How effective (and successful) is this new method?

REVISIT: THE FORMULATION

- Suppose:

- (x, y) : a test-time sample; $x \in R^d$ and $y \in [k]$; $x \in [0, 1]$
- f : a neural network; θ : its parameters
- $L(\theta, x, y)$: a loss function

- Goal (of the first order attacker):

- Find an $x^{adv} = x + \delta$ such that $\max_{\delta \in \mathcal{S}} L(\theta, x^{adv}, y)$ while $\|\delta\|_p \leq \varepsilon$

- PGD Crafts:

$$x^{t+1} = \Pi_{x+\mathcal{S}} \left(x^t + \alpha \operatorname{sgn}(\nabla_x L(\theta, x, y)) \right).$$

We Need to Know This!

OPTIMIZATION-BASED ATTACK IS THE GRADIENT ESTIMATION PROBLEM

- Zeroth-order Optimization
 - Finite Difference Method (FDM):

$$D_v f(x) = \langle \nabla_x f(x), v \rangle \approx (f(x + \delta v) - f(x)) / \delta.$$

- Compute: derivative of a function f at a point x towards a vector v

- FDM for the gradient with d -components:

$$\hat{\nabla}_x L(x, y) = \sum_{k=1}^d e_k (L(x + \delta e_k, y) - L(x, y)) / \delta \approx \sum_{k=1}^d e_k \langle \nabla_x L(x, y), e_k \rangle$$

- PGD in the black-box cases:

$$x^{t+1} = \Pi_{x+\mathcal{S}} (x^t + \alpha \operatorname{sgn}(\nabla_x L(\theta, x, y))).$$

OPTIMIZATION-BASED ATTACK IS THE GRADIENT ESTIMATION PROBLEM

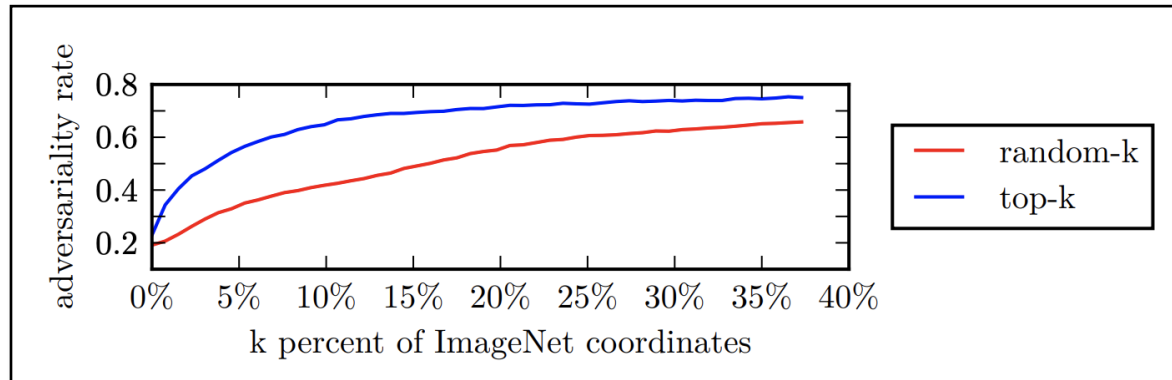
- Toy experiment

- Setup

- Compare the fraction of correctly estimated coordinates of gradients required
 - Compare top-k perturbations picked by magnitude or randomly
 - Measure the transfer-attack success rate

- Results:

- Adversarial attacks are effective even with the imperfect gradient estimate
 - Perturbations picked by magnitude is much effective than the random perturbations



OPTIMIZATION-BASED ATTACK IS THE GRADIENT ESTIMATION PROBLEM

- Prior approaches to do this estimation

- The Least Squares Method: $\min_{\hat{g}} \|\hat{g}\|_2 \quad \text{s.t. } A\hat{g} = y.$

- Iteratively compute the estimate \hat{g} , where:

- A : Queries $\{1, 2, \dots\}$
 - y : the corresponding inner product values

- Natural Evolution Strategy [Ilyas *et al.*] and Least Squares equivalence

$$\langle \hat{x}_{LSQ}, \mathbf{g} \rangle - \langle \hat{x}_{NES}, \mathbf{g} \rangle \leq O \left(\sqrt{\frac{k}{d} \cdot \log^3 \left(\frac{k}{p} \right)} \right) \|\mathbf{g}\|^2$$

OPTIMIZATION-BASED ATTACK IS THE GRADIENT ESTIMATION PROBLEM

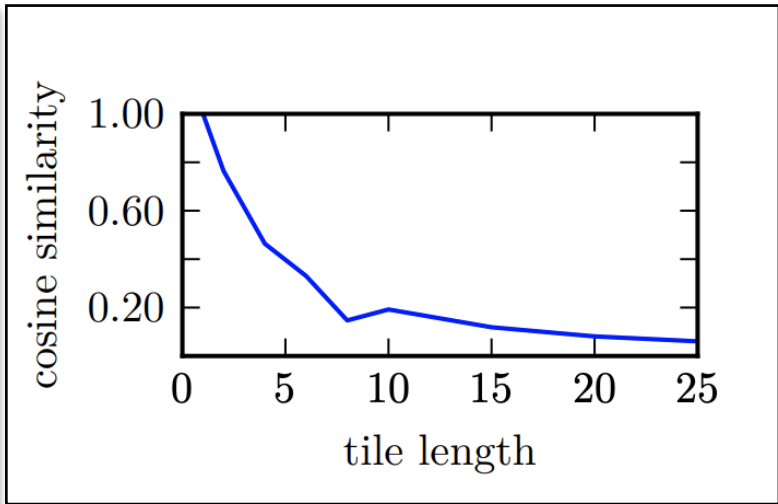
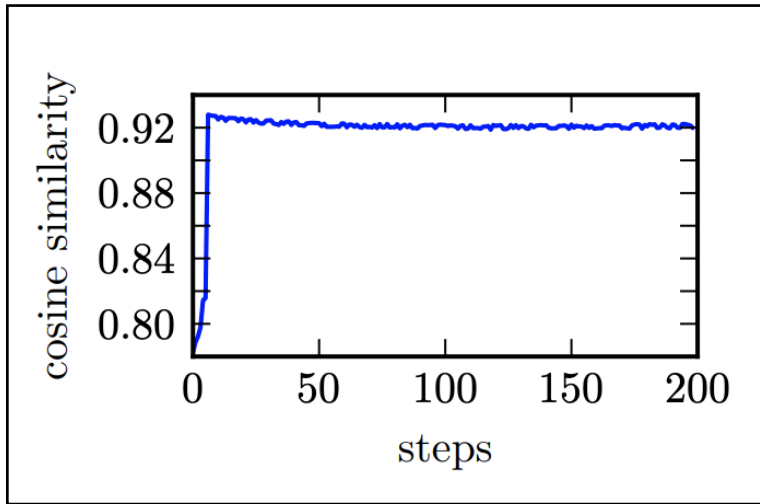
- Prior (= knowledge an adversary can acquire)
 - Gradients are correlated in successive attack iterations
 - Pixels close to each other tend to have similar values

OPTIMIZATION-BASED ATTACK IS THE GRADIENT ESTIMATION PROBLEM

- Prior (= knowledge an adversary can acquire)
 - [Time-dependent] Gradients are correlated in successive attack iterations
 - [Data-dependent] Pixels close to each other tend to have similar values

OPTIMIZATION-BASED ATTACK IS THE GRADIENT ESTIMATION PROBLEM

- Time-dependent & Data-dependent Priors



PUTTING ALL TOGETHER

- Formulate the Problem to the Bandit Framework
 - **Bandit problem**

Algorithm 1 Gradient Estimation with Bandit Optimization

```
1: procedure BANDIT-OPT-LOSS-GRAD-EST( $x, y_{init}$ )
2:    $v_0 \leftarrow \mathcal{A}(\phi)$ 
3:   for each round  $t = 1, \dots, T$  do
4:     // Our loss in round  $t$  is  $\ell_t(g_t) = -\langle \nabla_x L(x, y_{init}), g_t \rangle$ 
5:      $g_t \leftarrow v_{t-1}$ 
6:      $\Delta_t \leftarrow \text{GRAD-EST}(x, y_{init}, v_{t-1})$  // Estimated Gradient of  $\ell_t$ 
7:      $v_t \leftarrow \mathcal{A}(v_{t-1}, \Delta_t)$ 
8:    $g \leftarrow v_T$ 
9:   return  $\Pi_{\partial\mathcal{K}} [g]$ 
```

PUTTING ALL TOGETHER

- Formulate the Problem to the Bandit Framework
 - Gradient Estimation

Algorithm 2 Single-query spherical estimate of $\nabla_v \langle \nabla L(x, y), v \rangle$

```
1: procedure GRAD-EST( $x, y, v$ )
2:    $u \leftarrow \mathcal{N}(0, \frac{1}{\delta} I)$  // Query vector
3:    $\{q_1, q_2\} \leftarrow \{v + \delta u, v - \delta u\}$  // Antithetic samples
4:    $\ell_t(q_1) = -\langle \nabla L(x, y), q_1 \rangle \approx \frac{L(x, y) - L(x + \epsilon q_1, y)}{\epsilon}$  // Gradient estimation loss at  $q_1$ 
5:    $\ell_t(q_2) = -\langle \nabla L(x, y), q_2 \rangle \approx \frac{L(x, y) - L(x + \epsilon q_2, y)}{\epsilon}$  // Gradient estimation loss at  $q_2$ 
6:    $\Delta \leftarrow \frac{\ell_t(q_1) - \ell_t(q_2)}{\delta} u = \frac{L(x + \epsilon q_2, y) - L(x + \epsilon q_1, y)}{\delta \epsilon} u$ 
7:   // Note that due to cancellations we can actually evaluate  $\Delta$  with only two queries to  $L$ 
8:   return  $\Delta$ 
```

PUTTING ALL TOGETHER

- Formulate the Problem to the Bandit Framework
 - Gradient Estimation

Algorithm 3 Adversarial Example Generation with Bandit Optimization for ℓ_2 norm perturbations

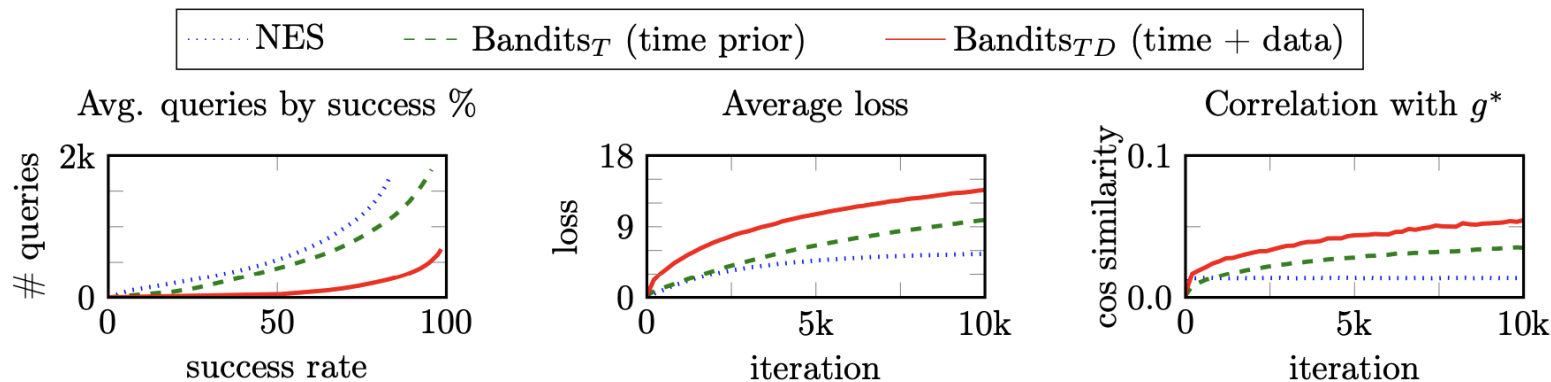
```
1: procedure ADVERSARIAL-BANDIT-L2( $x_{init}, y_{init}$ )
2:   //  $C(\cdot)$  returns top class
3:    $v_0 \leftarrow \mathbf{0}_{1 \times d}$  // If data prior,  $d < \dim(x)$ ;  $v_t$  ( $\Delta_t$ ) up (down)-sampled before (after) line 8
4:    $x_0 \leftarrow x_{init}$  // Adversarial image to be constructed
5:   while  $C(x) = y_{init}$  do
6:      $q_t \leftarrow v_{t-1}$ 
7:      $x_t \leftarrow x_{t-1} + h \cdot \frac{g_t}{\|g_t\|_2}$  // Boundary projection  $\frac{g}{\|g\|}$  standard PGD: c.f. [Rig15]
8:      $\Delta_t \leftarrow \text{GRAD-EST}(x_{t-1}, y_{init}, v_{t-1})$  // Estimated Gradient of  $\ell_t$ 
9:      $v_t \leftarrow v_{t-1} + \eta \cdot \Delta_t$ 
10:     $t \leftarrow t + 1$ 
return  $x_{t-1}$ 
```

HOW EFFECTIVE IS THIS NEW ATTACK (= METHOD)?

- Setup

- Dataset: ImageNet (10k randomly chosen samples)
- Model: Inception-v3
- Baseline: NES

- Results



OPTIMIZATION-BASED ATTACK

- Take aways
 - 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
 - How can we estimate gradient accurately with **smaller queries**?
 - Use two priors: time- and data-dependent priors
 - Formulate the estimation into the bandit framework
 - How **effective (and successful)** is this new method?
 - Require 2.5 – 5x less queries for successful attacks compared to NES

Thank You!

Instructor: Sanghyun Hong

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



Oregon State
University



TRUE AI
Trustworthy and Responsible AI