上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

SJTU

Learning Black-Box Attackers with Transferable Priors and Query Feedback

Jiancheng Yang*, Yangzhou Jiang*,

Xiaoyang Huang, Bingbing Ni, Chenglong Zhao.

Dec. 2020

NEURAL INFORMATION PROCESSING SYSTEMS

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

人工智能研究院
Artificial Intelligence Institute

Problem Setting

J. Yang, Y. Jiang, *et al.* NeurIPS 2020.

Black-box adversarial attack,
where only **classification confidence**
of a victim model is available.

It is a Siamese cat
(confidence: 99.9%).

Query

Black-Box
Victim Model

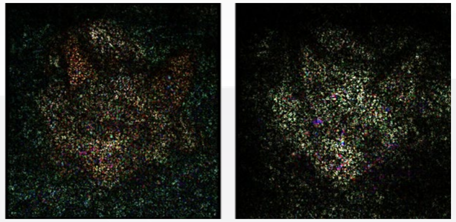
Attacker

How to make the victim
model make mistakes,
with **minimum queries**?

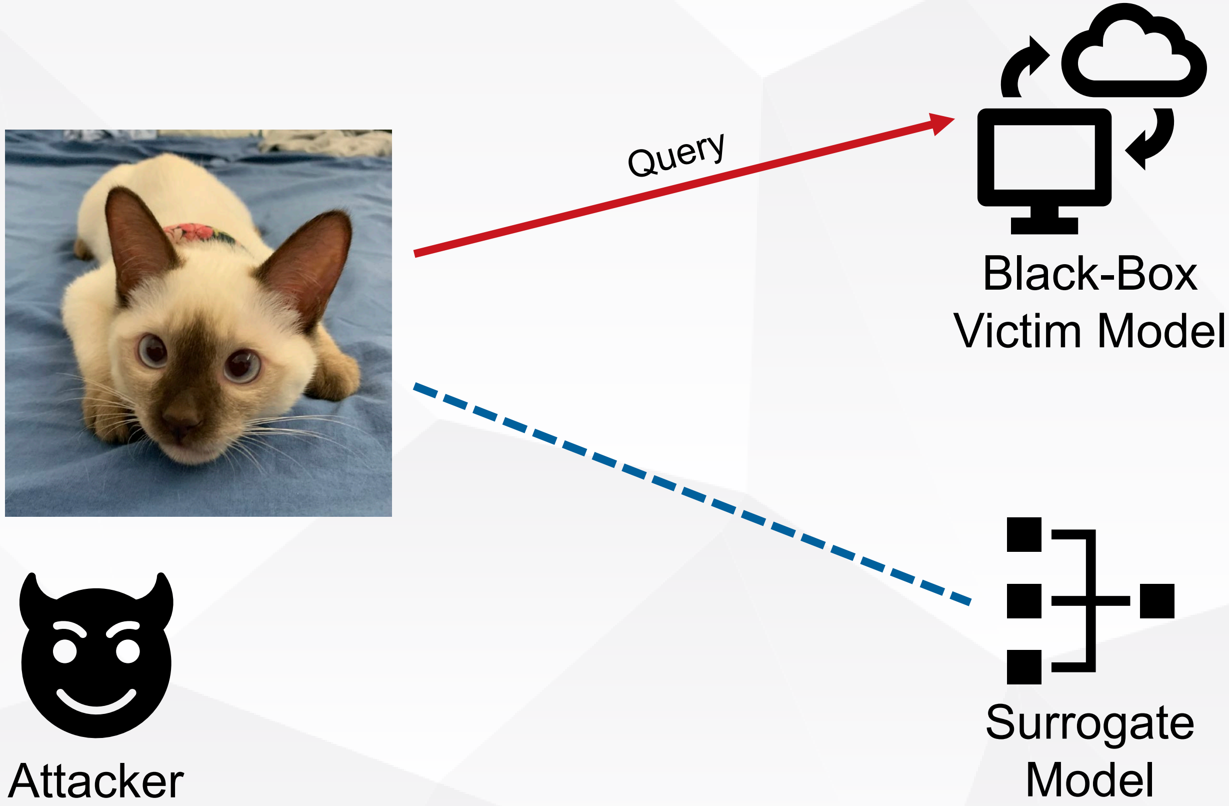
NEURAL INFORMATION
PROCESSING SYSTEMS

SJTU

Introducing a **surrogate model** to the victim model.



High consistency between gradients from vision models



Related work

J. Yang, Y. Jiang, *et al.* NeurIPS 2020.

Gradient Estimation Method

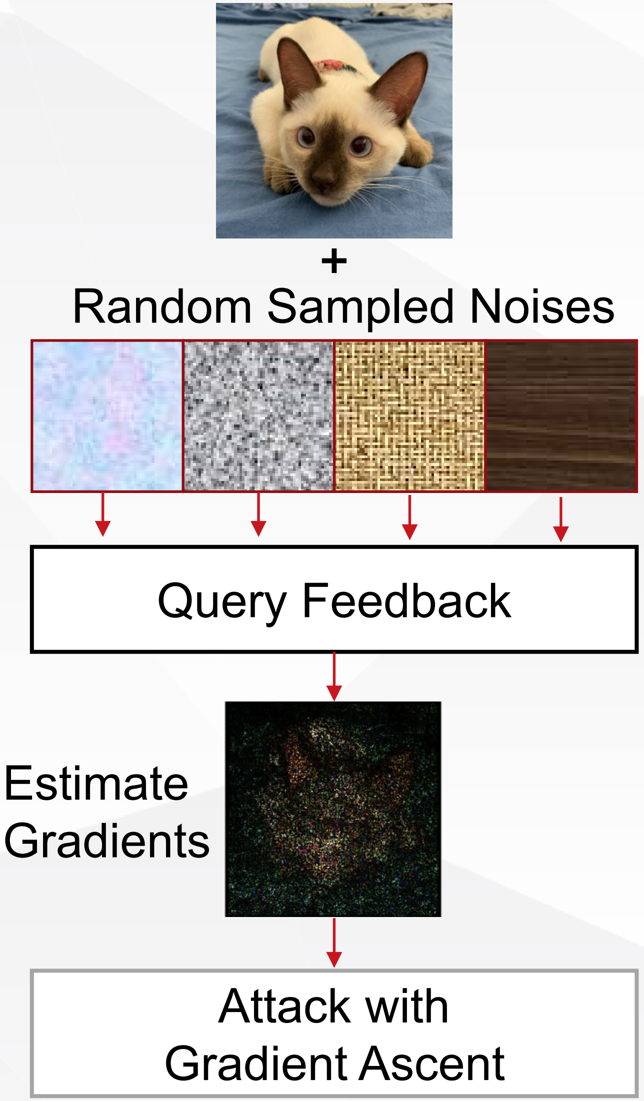
- Estimate gradient with **stochastic finite differences** (a.k.a **evolutionary strategies**)

$$\nabla \mathbb{E}[F(\theta)] \approx \frac{1}{\sigma n} \sum_{i=1}^{n} \delta_i F(\theta + \sigma \delta_i)$$

- Using **antithetic sampling** (using a pair of function evaluations at $(x + \epsilon \ and \ x - \epsilon)$ to reduce variance:

$$g = \frac{\beta}{2\sigma^2 P} \sum_{i=1}^{P} \epsilon_i \left(f(x + \epsilon_i) - f(x - \epsilon_i) \right)$$

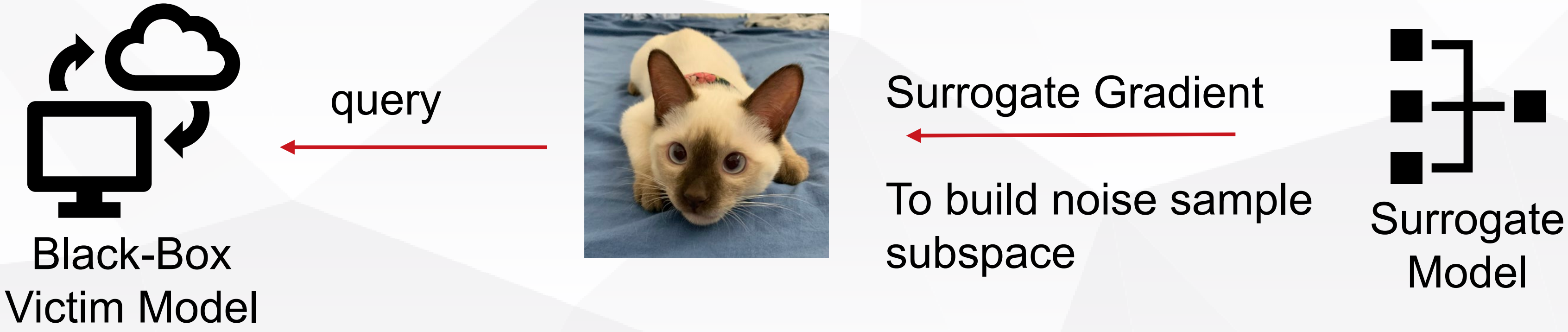
- Related methods:
  - NES (Natural Evolutionary Strategies) [1]
  - Bandit$_{TD}$ [2]
  - P-RGF$_{D}$ [3]



[1] Ilyas A, et al. Black-box adversarial attacks with limited queries and information. ICML'18.
[2] Ilyas A, et al. Prior convictions: Black-box adversarial attacks with bandits and priors. ICLR'19.
[3] Cheng S, et al. Improving black-box adversarial attacks with a transfer-based prior. NeurIPS'19.

Gradient Estimation with Surrogate Model

- Reduce **sampling space** with **surrogate gradient priors**.
- Related methods:
  - **PRGF$_D$** [1]:
    - Sample perturbation from surrogate gradient centered subspace.
    - Estimate optimal λ to balance between gradient prior and random search.
  - **Subspace attack** [2]:
    - Sample perturbation with gradients from a set of surrogate models.
    - Use dropout layer to obtain sample diversity.

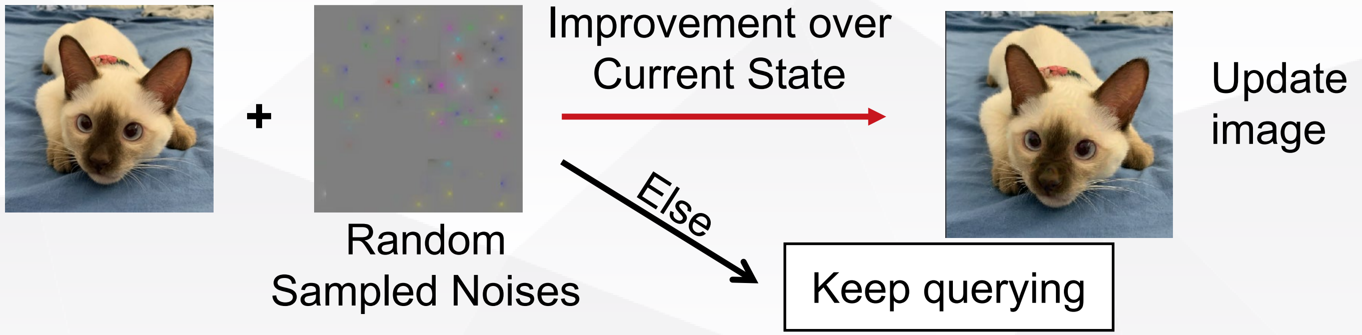


[1] Cheng S, et al. Improving black-box adversarial attacks with a transfer-based prior. NeurIPS'19.
[2] Guo Y et al. Subspace Attack: Exploiting Promising Subspaces for Query-Efficient Black-box Attacks. NeurIPS'19.

Related work

Random search

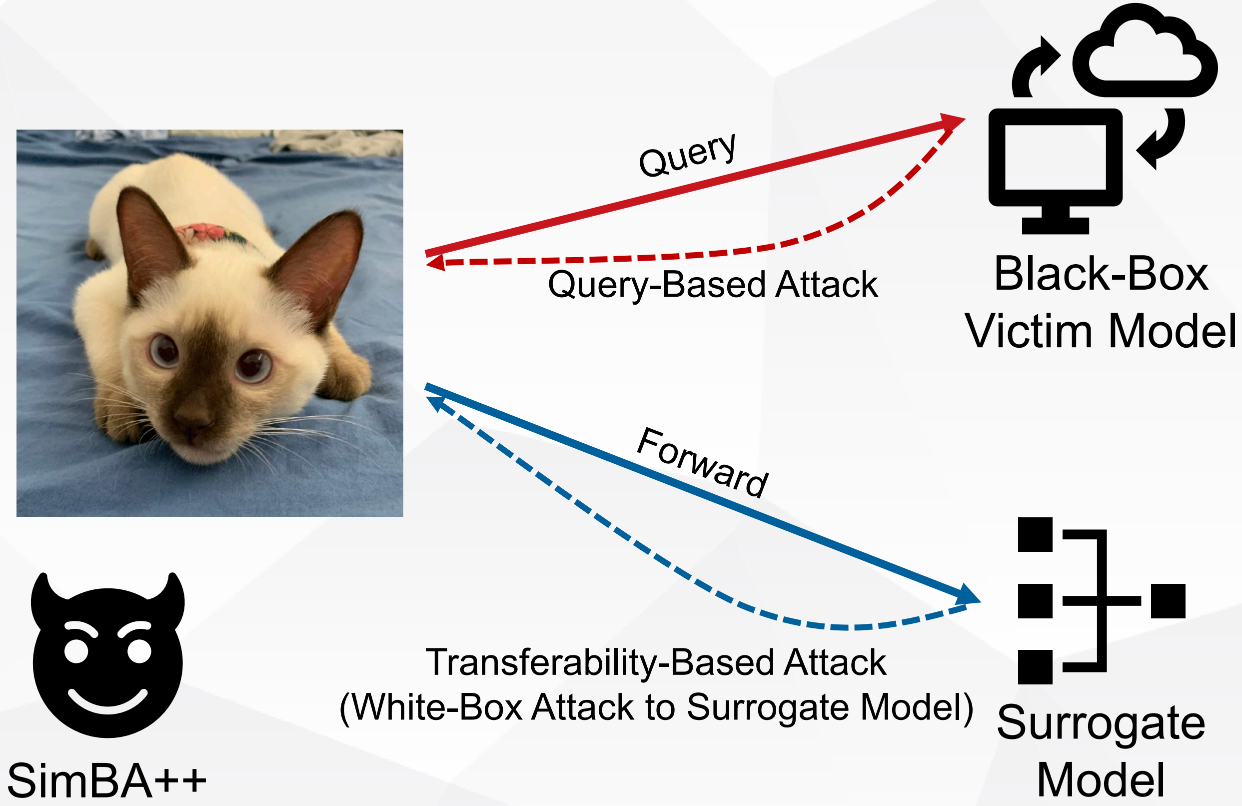
- Sample a random update δ at each iteration, and update greedily if it improves the objective function
- Related methods:
  - **SimBA / SimBA (DCT)** [1]
    - Randomly sample a vector from a predefined orthonormal basis in image space or frequency space, and either add or subtract it to the target image
  - **Square attack** [2]
    - Query target model with randomly sampled square-shaped noise.



[1] Guo C, et al. Simple black-box adversarial attacks. ICML'19.
[2] Andriushchenko M, et al. Square Attack: a query-efficient black-box adversarial attack via random search. ECCV'20.

Methodology: SimBA++

SimBA++: A strong baseline combining **transferability-based** and **query-based** black-box attack.



Please refer to the paper for a detailed algorithm.

Methodology: SimBA++

SimBA++: A strong baseline combining **transferability-based** and **query-based** black-box attack.

Pseudo Algorithm SimBA++

While not **Success** or **Exceed Attack Budget**:
    Every n_Q *iteration*:
        Run transferability-based attack (e.g., TIMI [1])
    Then:
        Run query-based attack (e.g., SimBA [2]) guided by surrogate model

Return adversarial example

This simple algorithm surprisingly **outperforms** several previous ***state of the art*** !

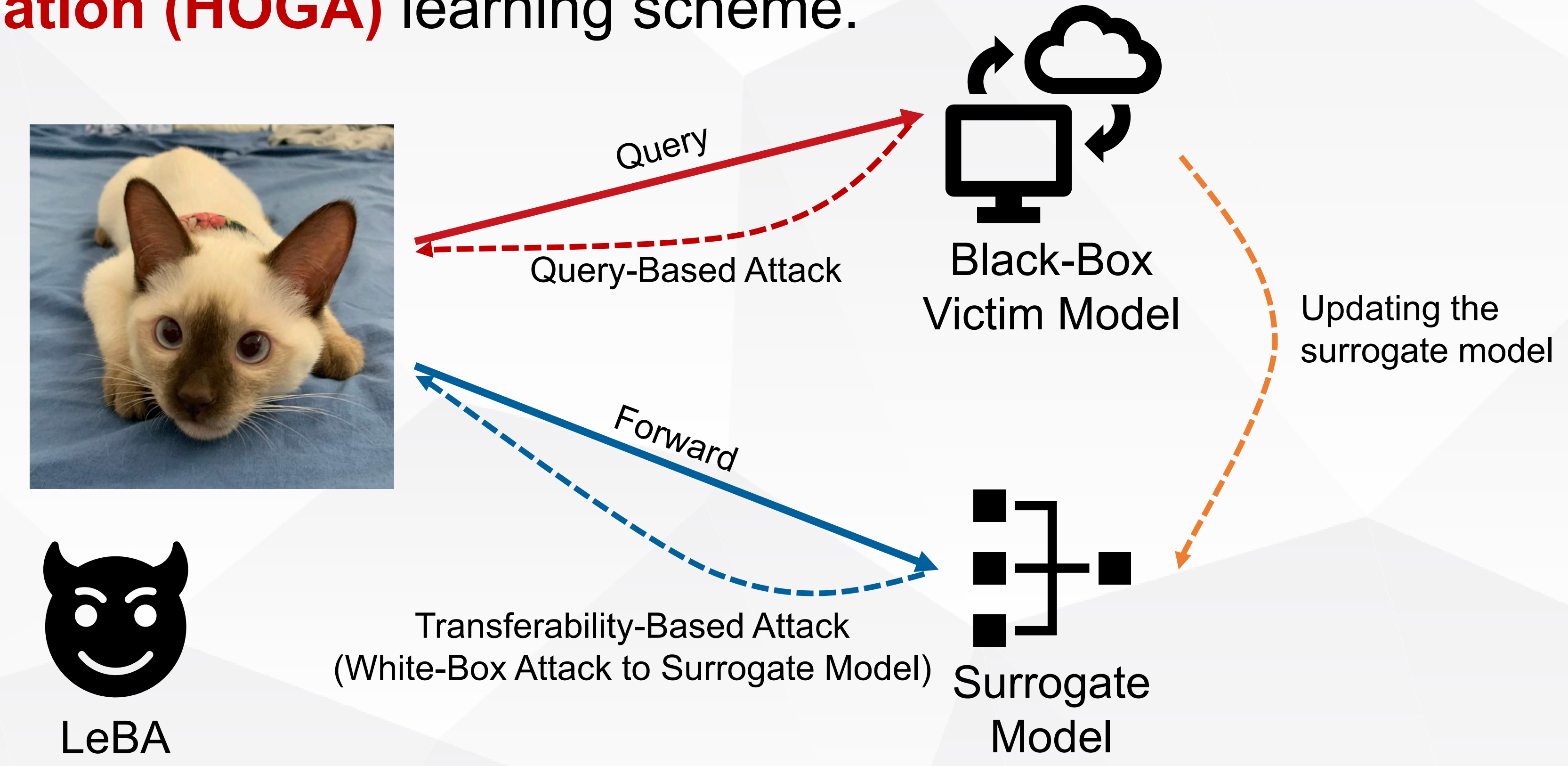
Please refer to the paper for a detailed algorithm.

NEURAL INFORMATION PROCESSING SYSTEMS

[1] Dong Y, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks. CVPR'19.
[2] Guo C, et al. Simple black-box adversarial attacks. ICML'19.

SJTU

Learnable Black-Box Attack (LeBA): Updating the surrogate model with **query feedback**, in a **High-Order Gradient Approximation (HOGA)** learning scheme.



Please refer to the paper for a detailed algorithm.

Learnable Black-Box Attack (LeBA): Updating the surrogate model with **query feedback**, in a **High-Order Gradient Approximation (HOGA)** learning scheme.

Pseudo Algorithm LeBA

While not **Success** or **Exceed Attack Budget**:
    Every n_Q *iteration*:
        Run transferability-based attack (e.g., TIMI [1])
    Then:
        Run query-based attack (e.g., SimBA [2]) guided by surrogate model
        Cache the query feedback
    Run **HOGA** to update the surrogate model
        to approximate **forward pass** and
        backward pass of victim model
Return adversarial example

Compute Forward Loss $l_F = MSE(\mathbf{S}_T, \mathbf{P}_T)$;
Create gradient graph and compute $\mathbf{g}_s = \frac{\partial log \mathbf{S}_T}{\partial \mathbf{X}_{adv}}$;
Compute Backward Loss l_B using
$l_B = MSE(\mathbf{g}_s(\mathbf{X}'_{adv} - \mathbf{X}_{adv}), \gamma(log\mathbf{P}'_T - log\mathbf{P}_T))$;
Back-propagate $l_B + \lambda l_F$ with high-order gradient;

NEURAL INFORMATION PROCESSING SYSTEMS

It improves the SimBA++ further!

Please refer to the paper for a detailed algorithm.

Illustration

Attack **success** with high **query efficiency** under l_2-norm threat model.

Clean: Siamese cat

Adversarial: Chihuahua

Attack Performance on ImageNet

J. Yang, Y. Jiang, *et al.* NeurIPS 2020.

High attack **success** rate (ASR) with improved **query efficiency**, even compared with recent Square Attack (ECCV'20).

Methods	Inception-V3		ResNet-50		VGG-16		Inception-V4		IncRes-V2	
	ASR	AVG.Q	ASR	AVG.Q	ASR	AVG.Q	ASR	AVG.Q	ASR	AVG.Q
NES [23] ICML'18	88.2%	1726.3	82.7%	1632.4	84.8%	1119.6	80.7%	2254.3	52.5%	3333.3
$\text{Bandits}_{\text{TD}}$ [24] ICLR'19	97.7%	836.1	93.0%	765.3	91.1%	275.9	96.2%	1170.9	89.7%	1569.3
Subspace [20] NeurIPS'19	96.6%	1635.8	94.4%	1078.7	96.2%	1085.8	94.7%	1838.2	91.2%	1780.6
RGF [10] NeurIPS'19	97.7%	1313.5	97.5%	1340.2	99.7%	823.2	93.2%	1860.1	85.6%	2135.3
P-RGF [10] NeurIPS'19	97.6%	750.8	98.7%	229.6	99.9%	685.5	96.5%	1095.6	88.9%	1380.2
P-RGF_{D} [10] NeurIPS'19	99.0%	637.4	99.3%	270.5	99.8%	393.1	98.3%	913.6	93.6%	1364.5
Square [2] ECCV'20	**99.4%**	351.9	99.8%	401.4	**100.0%**	**142.3**	98.3%	475.6	94.9%	670.3
TIMI [14] CVPR'19	49.0%	-	68.6%	-	51.3%	-	44.3%	-	44.5%	-
SimBA [19] ICML'19	97.8%	874.5	99.6%	873.9	**100.0%**	423.3	96.2%	1149.8	92.0%	1516.1
SimBA+ (Ours)	98.2%	725.2	99.7%	717.0	**100.0%**	365.9	96.8%	946.2	92.5%	1234.7
SimBA++ (Ours)	99.2%	295.7	**99.9%**	187.3	99.9%	166.0	98.3%	420.2	95.8%	555.1
LeBA (Ours)	**99.4%**	**243.8**	**99.9%**	**178.7**	99.9%	145.5	**98.7%**	**347.4**	**96.6%**	**514.2**

NEURAL INFORMATION PROCESSING SYSTEMS

SJTU

Attack over Defensive Methods

J. Yang, Y. Jiang, *et al.* NeurIPS 2020.

High attack **success** rate (ASR) with improved **query efficiency**, even compared with recent Square Attack (ECCV'20).

Methods	JPEG Compression ASR	JPEG Compression AVG.Q	Guided Denoiser ASR	Guided Denoiser AVG.Q	Adversarial Training ASR	Adversarial Training AVG.Q
NES [23] ICML'18	14.9%	2330.9	57.6%	2773.8	59.4%	2773.6
$\text{Bandits}_{\text{TD}}$ [24] ICLR'19	95.8%	1086.7	20.3%	759.6	96.6%	1121.4
Subspace [20] NeurIPS'19	46.7%	2073.4	93.2%	1619.2	93.4%	1651.7
RGF [10] NeurIPS'19	74.4%	846.9	22.0%	2419.1	87.6%	2095.3
P-RGF_{D} [10] NeurIPS'19	94.8%	751.2	82.6%	1588.3	98.4%	1092.8
Square [2] ECCV'20	**98.8%**	342.3	98.2%	392.6	98.5%	387.6
TIMI [14] CVPR'19	48.2%	-	39.3%	-	39.2%	-
SimBA [19] ICML'19	96.0%	762.8	98.0%	971.6	98.0%	978.0
SimBA+ (Ours)	96.8%	663.4	98.2%	797.1	98.0%	779.4
SimBA++ (Ours)	98.2%	325.1	98.5%	407.9	98.7%	422.9
LeBA (Ours)	**98.8%**	**273.0**	**98.8%**	**343.6**	**98.9%**	**355.0**

NEURAL INFORMATION PROCESSING SYSTEMS

Updating the Surrogate Model

The updated surrogate model trained on Data S1

- works **better** than original surrogated model: LeBA (*test*) > LeBA (*training*)
- could **generalize** to **new Data** S2: LeBA (*test*) > SimBA++

		Inception-V3		ResNet-50		VGG-16		Inception-V4		IncRes-V2	
Data	Methods	ASR	AVG.Q	ASR	AVG.Q	ASR	AVG.Q	ASR	AVG.Q	ASR	AVG.Q
S1	SimBA++	99.2%	295.7	**99.9%**	187.3	**99.9%**	166.0	98.3%	420.2	95.8%	555.1
	LeBA (*training*)	**99.4%**	243.8	**99.9%**	178.7	**99.9%**	145.5	**98.7%**	347.4	**96.6%**	514.2
	LeBA (*test*)	**99.4%**	**230.6**	**99.9%**	**172.3**	**99.9%**	**138.5**	98.4%	**322.4**	**96.6%**	**510.2**
S2	SimBA++	99.7%	183.0	**100.0%**	110.4	**100.0%**	98.6	98.8%	245.1	**97.6%**	325.8
	LeBA (*test*)	**99.8%**	**151.3**	**100.0%**	**97.2**	**100.0%**	**96.2**	**98.9%**	**215.9**	**97.6%**	**290.8**

NEURAL INFORMATION PROCESSING SYSTEMS

SJTU

Conclusion

- We propose **SimBA++** and **Learnable Black-Box Attack (LeBA)** by combing transferability-based and query-based attack.
- With a novel **High-Order Gradient Approximation (HOGA)** scheme, we update the surrogate model within limited queries.
- The proposed methods empirically establish a new ***state of the art***, in terms of attack success and query efficiency.

Check out the code for this study

https://github.com/TrustworthyDL/LeBA

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks for Listening

WeChat
jekyll4168

WeChat
j1098122556

饮 水 思 源　爱 国 荣 校