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

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

Query

Black-Box
Victim Model

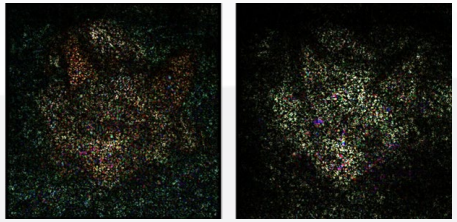
It is a Siamese cat
(confidence: 99.9%).

Attacker

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

Motivation

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



High consistency between gradients from vision models

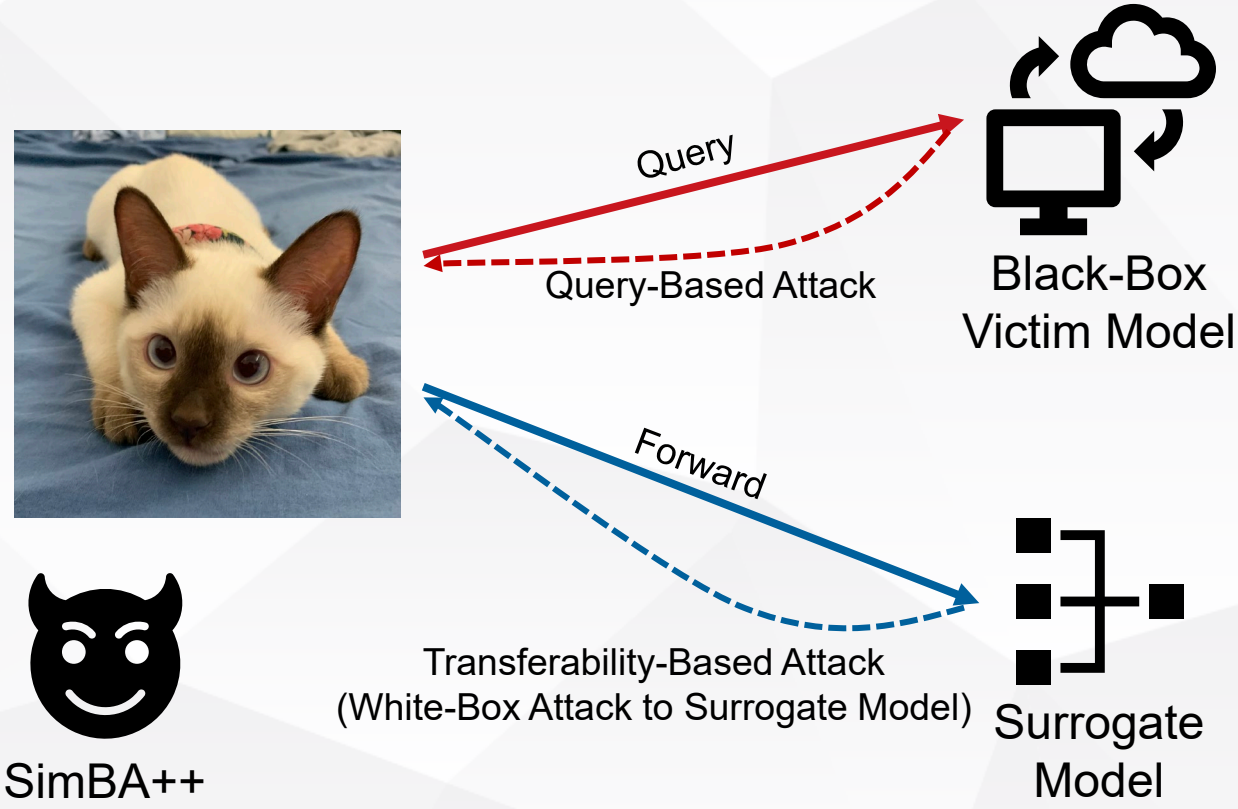
Query

Black-Box Victim Model

Attacker

Surrogate Model

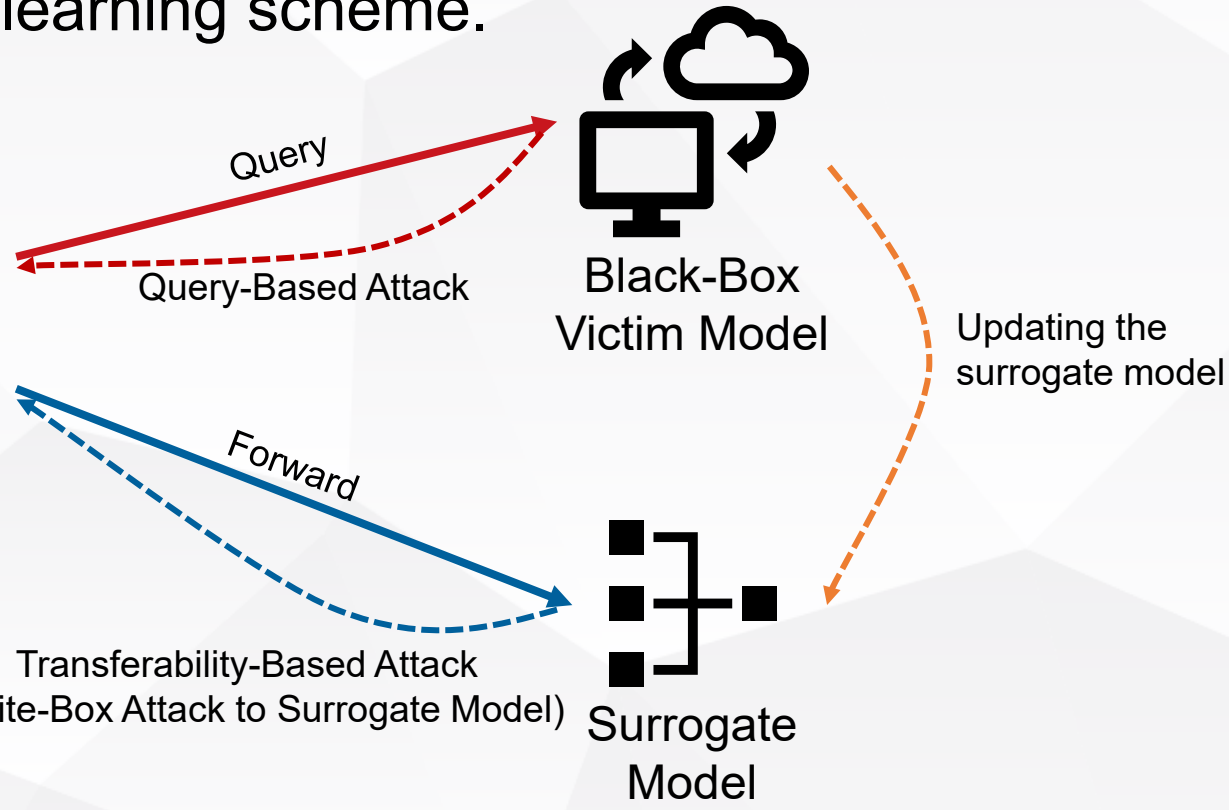
SimBA++: A strong baseline combining **transferability-based** and **query-based** black-box attack.



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.

LeBA



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**

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		Guided Denoiser		Adversarial Training	
	ASR	AVG.Q	ASR	AVG.Q	ASR	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

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

Thanks for Listening

WeChat
jekyll4168

WeChat
j1098122556

饮 水 思 源　　爱 国 荣 校