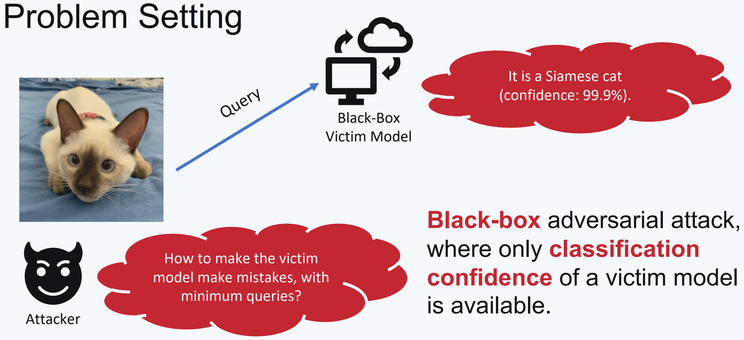
Learning Black-Box Attackers with Transferable Priors and Query Feedback

Jiancheng Yang*, Yangzhou Jiang*, Xiaoyang Huang, Bingbing Ni, Chenglong Zhao

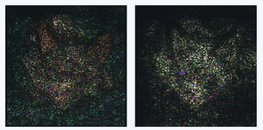
NEURAL INFORMATION PROCESSING SYSTEMS

Problem Setting



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

Motivation



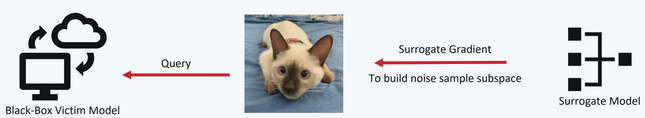
High consistency between gradients from vision models

- Introducing a **surrogate model.**
- Leverage the **transferability-based** (low cost, low success) and **query-based** (high cost, high cost) attack.
- Use the **query feedback** to **update** the surrogate model.

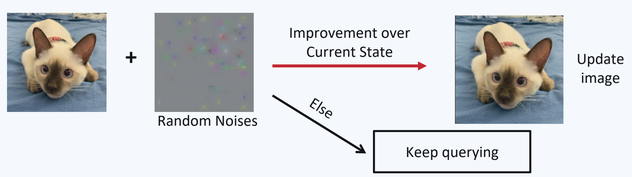
Related Work

Gradient Estimation: NES [1], Bandit$_{TD}$ [2], RGF [3]

Gradient Estimation with Surrogate Model: P-RGF [3], Subspace [4]

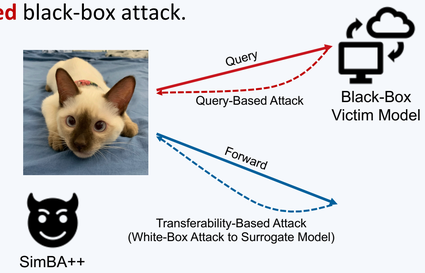


Random Search: SimBA [5], Square [6]



Methodology (Random Search + Surrogate Model)

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



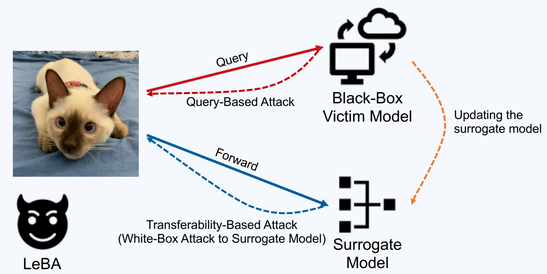
Pseudo Algorithm SimBA++

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*** !

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

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;

Return adversarial example

It improves SimBA++ further!

Results

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

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

Methods	Inception-V3 ASR	Inception-V3 AVG.Q	ResNet-50 ASR	ResNet-50 AVG.Q	VGG-16 ASR	VGG-16 AVG.Q	Inception-V4 ASR	Inception-V4 AVG.Q	IncRes-V2 ASR	IncRes-V2 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
Bandits$_{TD}$ [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**

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
Bandits$_{TD}$ [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**

References

[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
[4] Guo Y et al. Subspace Attack: Exploiting Promising Subspaces for Query-Efficient Black-box Attacks. NeurIPS'19.
[5] Guo C, et al. Simple black-box adversarial attacks. ICML'19.
[6] Andriushchenko M, et al. Square Attack: a query-efficient black-box adversarial attack via random search. ECCV'20.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Open-Source