Tsinghua University

CVPR VIRTUAL JUNE 19-25

Simulating Unknown Target Models for Query-Efficient Black-box Attacks

Chen Ma, Li Chen, Jun-Hai Yong

School of Software, BNRist, Tsinghua University

Motivation

Challenge: How to reduce the high query complexity of black-box attack remains an open problem.

Model stealing attacks can replicate the functionality of target model.

How about counterfeit the target model by using a substitute model to **transfer the query stress**?

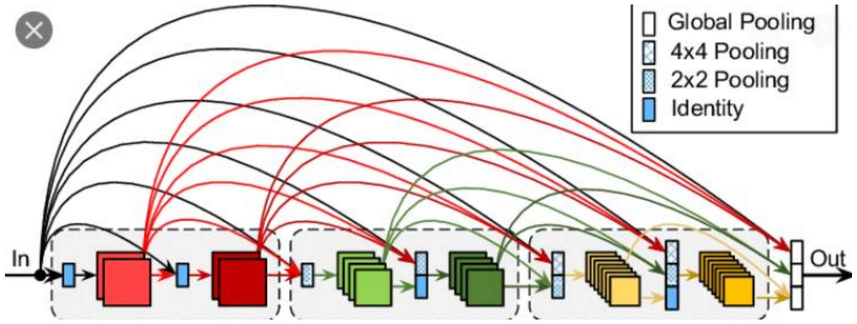
However, the training requires querying the target model. Consequently, the query complexity remains high, and such attacks can be defended easily.

Motivation

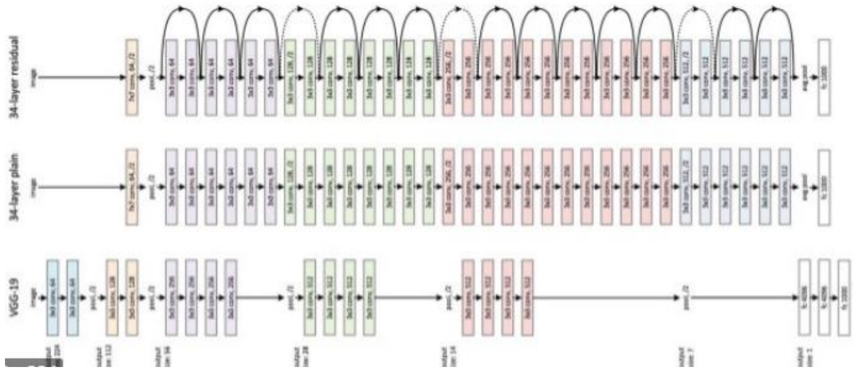
How to train a substitute model without the target model requirement is worthy of further exploration.

One network simulates them all!

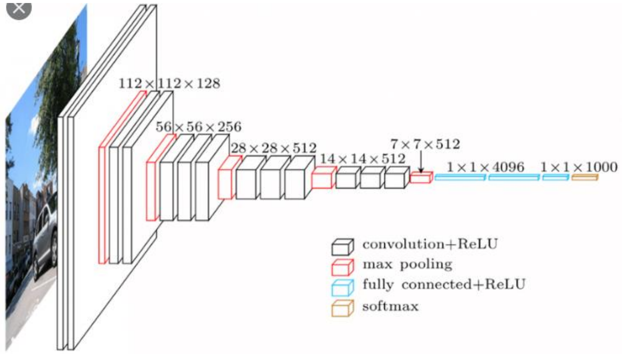
Simulator



DenseNet



ResNet-101

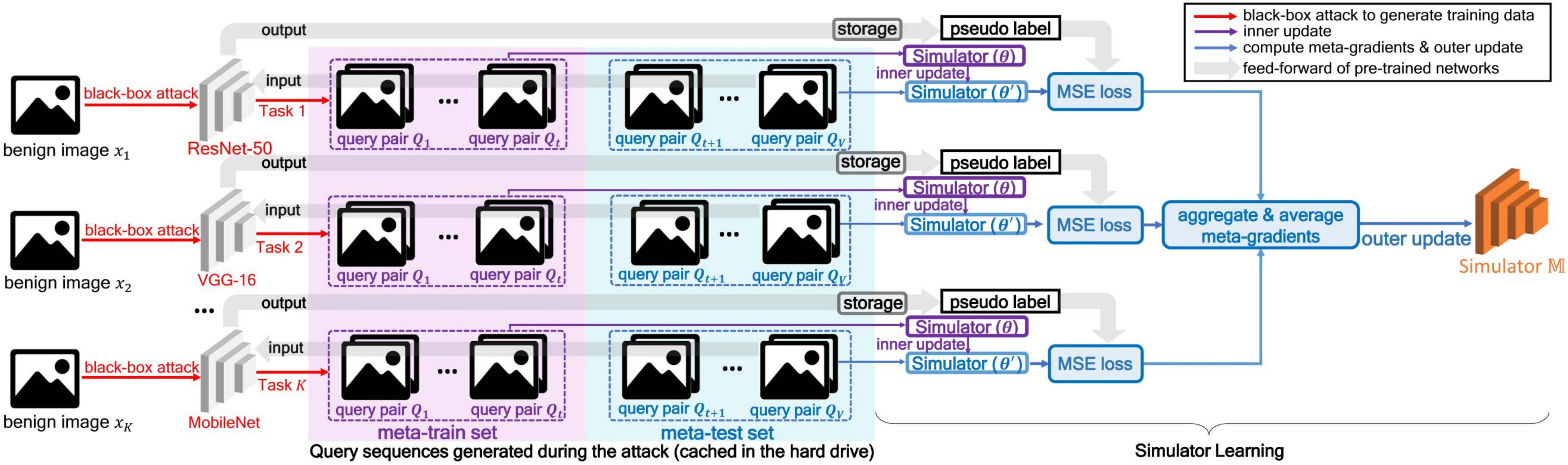


VGG-16

Method

Two stages: **meta-training** and **attack stage.**

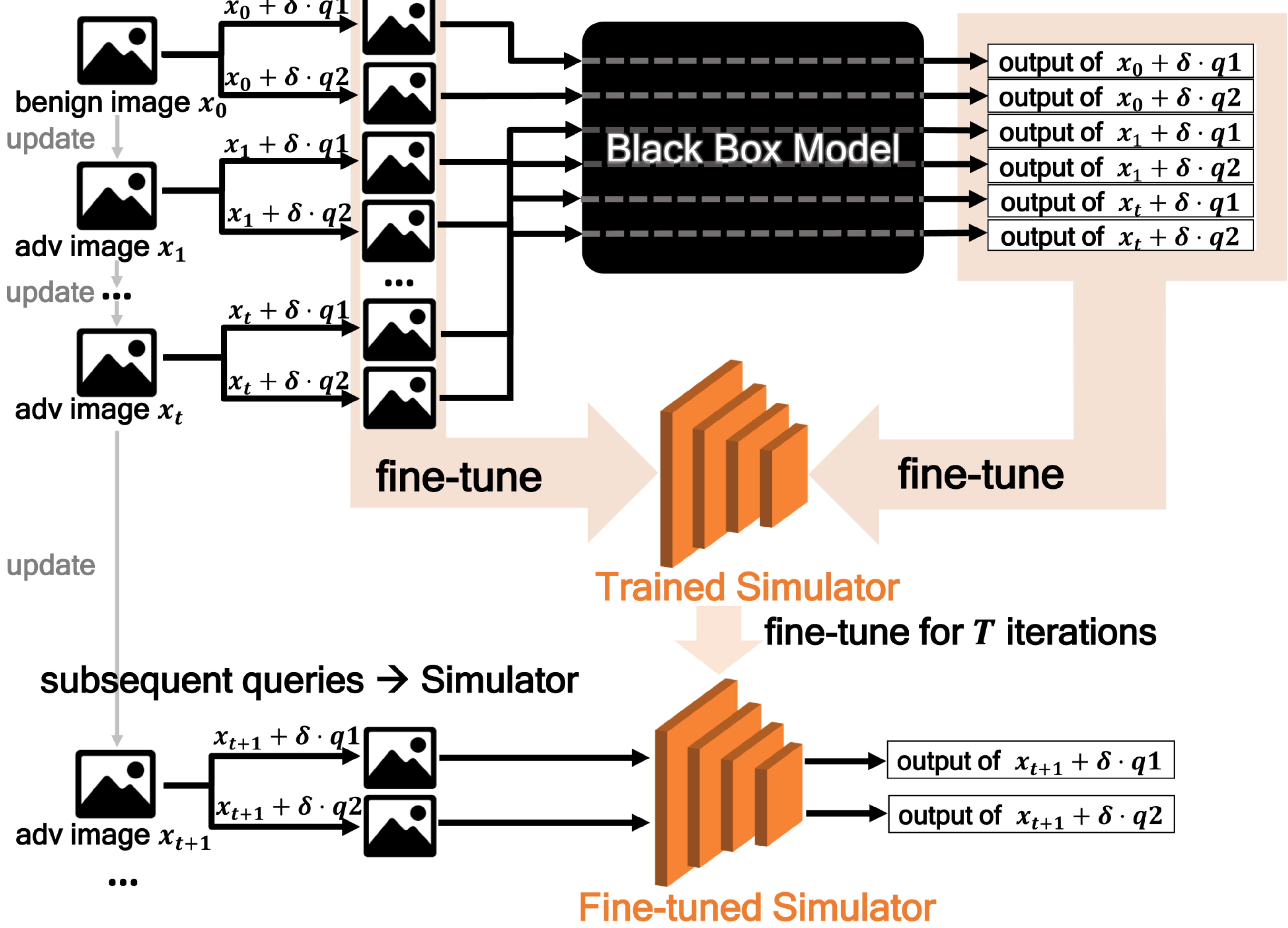
Training: we collect intermediate query sequences generated by attacking different existing networks. Each sequence is divided into meta-train and meta-test.



M

Attack ator and

the targ



Algorithm 1 Training procedure of the Simulator

Input: Training dataset D, Bandits attack algorithm $\mathcal{A}$, pre-trained classification networks $\mathbb{N}_1, \ldots, \mathbb{N}_n$, the Simulator network $\mathbb{M}$ and its parameters θ, feed-forward function f of $\mathbb{M}$, loss function $\mathcal{L}(\cdot,\cdot)$ defined in Eq. (1).
Parameters: Training iterations N, query sequence size V, meta-train set size t, batch size K, inner-update learning rate λ_1, outer-update learning rate λ_2, inner-update iterations T.
Output: The learned Simulator $\mathbb{M}$.

for $iter \leftarrow 1$ to N **do**
  sample K benign images $x_1, \ldots, x_K$ from D
  for $k \leftarrow 1$ to K **do** ▷ iterate over K tasks
    a network $\mathbb{N}_i \leftarrow$ sample from $\mathbb{N}_1, \ldots, \mathbb{N}_n$
    $Q_1, \ldots, Q_V \leftarrow \mathcal{A}(x_k, \mathbb{N}_i)$ ▷ query sequence
    $\mathcal{D}_{mtr} \leftarrow Q_1, \ldots, Q_t$
    $\mathcal{D}_{mte} \leftarrow Q_{t+1}, \ldots, Q_V$
    $\mathbf{p}_{\texttt{train}} \leftarrow \mathbb{N}_i(\mathcal{D}_{mtr})$
    $\mathbf{p}_{\texttt{test}} \leftarrow \mathbb{N}_i(\mathcal{D}_{mte})$ ▷ pseudo labels
    $\theta' \leftarrow \theta$ ▷ reinitialize $\mathbb{M}$'s weights
    for $j \leftarrow 1$ to T **do**
      $\theta' \leftarrow \theta' - \lambda_1 \cdot \nabla_{\theta'} \mathcal{L}\left(f_{\theta'}\left(\mathcal{D}_{mtr}\right), \mathbf{p}_{\texttt{train}}\right)$
    end for
    $L_i \leftarrow \mathcal{L}\left(f_{\theta'}\left(\mathcal{D}_{mte}\right), \mathbf{p}_{\texttt{test}}\right)$
  end for
  $\theta \leftarrow \theta - \lambda_2 \cdot \frac{1}{K} \sum_{i=1}^{K} \nabla_\theta L_i$ ▷ the outer update
end for
return $\mathbb{M}$

Algorithm 2 Simulator Attack under the ℓ_p norm constraint

Input: Input image $x \in \mathbb{R}^D$ where D is the image dimensionality, true label y of x, feed-forward function f of target model, Simulator $\mathbb{M}$, attack objective loss $\mathcal{L}(\cdot,\cdot)$.
Parameters: Warm-up iterations t, simulator-predict interval m, Bandits exploration τ, finite difference probe δ, OCO learning rate η_g, image learning rate η.
Output: x_{adv} that satisfies $\|x_{\text{adv}} - x\|_p \le \epsilon$.

Initialize the adversarial example $x_{\text{adv}} \leftarrow x$
Initialize the gradient to be estimated $\mathbf{g} \leftarrow \mathbf{0}$
Initialize $\mathbb{D} \leftarrow deque(maxlen = t)$ ▷ a bounded double-ended queue with maximum length of t, adding a full $\mathbb{D}$ leads it to drop its oldest item automatically.
for $i \leftarrow 1$ to N **do**
  $\mathbf{u} \leftarrow \mathcal{N}(\mathbf{0}, \frac{1}{D}\mathbf{I})$ ▷ the same dimension with x
  $q1 \leftarrow \mathbf{g} + \tau\mathbf{u}, \quad q2 \leftarrow \mathbf{g} - \tau\mathbf{u}$
  $q1 \leftarrow q1/\|q1\|_2, \quad q2 \leftarrow q2/\|q2\|_2$
  if $i \le t$ **or** $(i - t)$ mod $m = 0$ **then**
    $\hat{y}_1 \leftarrow f(x_{\text{adv}} + \delta \cdot q1)$
    $\hat{y}_2 \leftarrow f(x_{\text{adv}} + \delta \cdot q2)$
    $\{x_{\text{adv}} + \delta \cdot q1, \hat{y}_1, x_{\text{adv}} + \delta \cdot q2, \hat{y}_2\}$ append $\mathbb{D}$
    if $i \ge t$ **then**
      Fine-tune $\mathbb{M}$ using $\mathbb{D}$ ▷ fine-tune $\mathbb{M}$ every m iterations after the warm-up phase.
    end if
  else
    $\hat{y}_1 \leftarrow \mathbb{M}(x_{\text{adv}} + \delta \cdot q1), \quad \hat{y}_2 \leftarrow \mathbb{M}(x_{\text{adv}} + \delta \cdot q2)$
  end if
  $\Delta_g \leftarrow \frac{\mathcal{L}(\hat{y}_1, y) - \mathcal{L}(\hat{y}_2, y)}{\tau\delta}\mathbf{u}$
  if $p = 2$ **then**
    $\mathbf{g} \leftarrow \mathbf{g} + \eta_g \cdot \Delta_g$
    $x_{\text{adv}} \leftarrow \prod_{\mathcal{B}_2(x,\epsilon)}(x_{\text{adv}} + \eta \cdot \frac{\mathbf{g}}{\|\mathbf{g}\|_2})$ ▷ $\prod_{\mathcal{B}_p(x,\epsilon)}$ denotes the ℓ_p norm projection under ℓ_p norm bound.
  else if $p = \infty$ **then** ▷ using the exponentiated gradient update [20] in the ℓ_∞ norm attack as follows.
    $\hat{\mathbf{g}} \leftarrow \frac{\mathbf{g}+1}{2}$
    $\mathbf{g} \leftarrow \frac{\hat{\mathbf{g}} \cdot \exp(\eta_g \cdot \Delta_g) - (\mathbf{1} - \hat{\mathbf{g}}) \cdot \exp(-\eta_g \cdot \Delta_g)}{\hat{\mathbf{g}} \cdot \exp(\eta_g \cdot \Delta_g) + (\mathbf{1} - \hat{\mathbf{g}}) \cdot \exp(-\eta_g \cdot \Delta_g)}$
    $x_{\text{adv}} \leftarrow \prod_{\mathcal{B}_\infty(x,\epsilon)}(x_{\text{adv}} + \eta \cdot \text{sign}(\mathbf{g}))$
  end if
  $x_{\text{adv}} \leftarrow \text{Clip}(x_{\text{adv}}, 0, 1)$
end for
return x_{adv}

Results

Untargeted Attack in CIFAR-10/CIFAR-100

Dataset	Norm	Attack	Attack Success Rate				Avg. Query				Median Query			
			PyramidNet-272	GDAS	WRN-28	WRN-40	PyramidNet-272	GDAS	WRN-28	WRN-40	PyramidNet-272	GDAS	WRN-28	WRN-40
CIFAR-10	ℓ_2	NES [19]	99.5%	74.8%	99.9%	99.5%	200	123	159	154	150	100	100	100
		RGF [31]	100%	100%	100%	100%	216	168	153	150	204	152	102	152
		P-RGF [8]	100%	100%	100%	100%	**64**	40	76	73	62	**20**	64	64
		Meta Attack [12]	99.2%	99.4%	98.6%	99.6%	2359	1611	1853	1707	2211	1303	1432	1430
		Bandits [20]	100%	100%	100%	100%	151	66	107	98	110	54	80	78
		Simulator Attack	100%	100%	100%	100%	92	**34**	**48**	**51**	**52**	26	**34**	**34**
	ℓ_∞	NES [19]	86.8%	71.4%	74.2%	77.5%	1559	628	1235	1209	600	300	400	400
		RGF [31]	99%	93.8%	98.6%	98.8%	955	646	1178	928	668	460	663	612
		P-RGF [8]	97.3%	97.9%	97.7%	98%	**742**	337	703	564	**408**	128	236	217
		Meta Attack [12]	90.6%	98.8%	92.7%	94.2%	3456	2034	2198	1987	2991	1694	1564	1433
		Bandits [20]	99.6%	100%	99.4%	99.9%	1015	391	611	542	560	166	224	228
		Simulator Attack	96.5%	99.9%	98.1%	98.8%	779	**248**	**466**	**419**	469	**83**	**186**	**186**
CIFAR-100	ℓ_2	NES [19]	92.4%	90.2%	98.4%	99.6%	118	94	102	105	100	50	100	100
		RGF [31]	100%	100%	100%	100%	114	110	106	106	102	101	102	102
		P-RGF [8]	100%	100%	100%	100%	54	46	54	73	62	62	62	62
		Meta Attack [12]	99.7%	99.8%	99.4%	98.4%	1022	930	1193	1252	783	781	912	913
		Bandits [20]	100%	100%	100%	100%	58	54	64	65	42	42	52	53
		Simulator Attack	100%	100%	100%	100%	**29**	**29**	**33**	**34**	**24**	**24**	**26**	**26**
	ℓ_∞	NES [19]	91.3%	89.7%	92.4%	89.3%	439	271	673	596	204	153	255	255
		RGF [31]	99.7%	98.8%	98.9%	98.9%	385	420	544	619	256	255	357	357
		P-RGF [8]	99.3%	98.2%	98%	97.8%	308	220	371	480	147	116	136	181
		Meta Attack [12]	99.7%	99.8%	97.4%	97.3%	1102	1098	1294	1369	912	911	1042	1040
		Bandits [20]	100%	100%	99.8%	99.8%	266	209	262	260	68	57	107	92
		Simulator Attack	100%	100%	99.9%	99.9%	**129**	**124**	**196**	**209**	**34**	**28**	**58**	**54**

Results

Targeted Attack in CIFAR-10/CIFAR-100

Dataset	Norm	Attack	Attack Success Rate				Avg. Query				Median Query			
			PyramidNet-272	GDAS	WRN-28	WRN-40	PyramidNet-272	GDAS	WRN-28	WRN-40	PyramidNet-272	GDAS	WRN-28	WRN-40
CIFAR-10	ℓ_2	NES [19]	93.7%	95.4%	98.5%	97.7%	1474	1515	1043	1088	1251	999	881	882
		Meta Attack [12]	92.2%	97.2%	74.1%	74.7%	4215	3137	3996	3797	3842	2817	3586	3329
		Bandits [20]	99.7%	100%	97.3%	98.4%	852	718	1082	997	458	538	338	399
		Simulator Attack (m=3)	99.1%	100%	98.5%	95.6%	896	718	990	980	373	**388**	217	249
		Simulator Attack (m=5)	97.6%	99.9%	96.4%	94%	**815**	**715**	**836**	**793**	**368**	400	**206**	**245**
	ℓ_∞	NES [19]	63.8%	80.8%	89.7%	88.8%	4355	3942	3046	3051	3717	3441	2535	2592
		Meta Attack [12]	75.6%	95.5%	59%	59.8%	4960	3461	3873	3899	4736	3073	3328	3586
		Bandits [20]	84.5%	98.3%	76.9%	79.8%	2830	1755	2037	2128	2081	1162	1178	1188
		Simulator Attack (m=3)	80.9%	97.8%	83.1%	82.2%	2655	1561	1855	1806	1943	918	1010	1018
		Simulator Attack (m=5)	78.7%	96.5%	80.8%	80.3%	**2474**	**1470**	**1676**	**1660**	**1910**	**917**	**957**	**956**
CIFAR-100	ℓ_2	NES [19]	87.6%	77%	89.3%	87.6%	1300	1405	1383	1424	1102	1172	1061	1049
		Meta Attack [12]	86.1%	88.7%	63.4%	43.3%	4000	3672	4879	4989	3457	3201	4482	4865
		Bandits [20]	99.6%	100%	98.9%	91.5%	1442	847	1645	2436	1058	679	1150	1584
		Simulator Attack (m=3)	99.3%	100%	98.6%	92.6%	921	724	1150	1552	666	519	779	1126
		Simulator Attack (m=5)	97.8%	99.6%	95.7%	83.9%	**829**	**679**	**1000**	**1211**	**644**	**508**	**706**	**906**
	ℓ_∞	NES [19]	72.1%	66.8%	68.4%	69.9%	4673	5174	4763	4770	4376	4832	4357	4508
		Meta Attack [12]	80.4%	81.2%	57.6%	40.1%	4136	3951	4893	4967	3714	3585	4609	4737
		Bandits [20]	81.2%	92.5%	72.4%	56%	3222	2798	3353	3465	2633	2132	2766	2774
		Simulator Attack (m=3)	89.4%	94.2%	79%	64.3%	2732	2281	3078	3238	1854	1589	2185	2548
		Simulator Attack (m=5)	83.7%	91.4%	74.2%	60%	**2410**	**2134**	**2619**	**2823**	**1754**	**1572**	**2080**	**2270**

Results

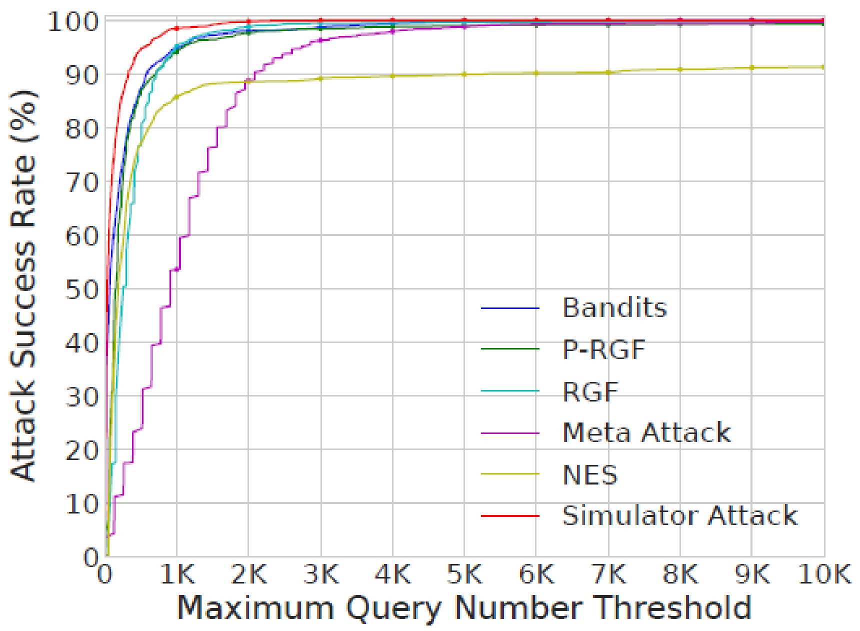
Attack	Attack Success Rate			Avg. Query			Median Query		
	D_{121}	R_{32}	R_{64}	D_{121}	R_{32}	R_{64}	D_{121}	R_{32}	R_{64}
NES [19]	74.3%	45.3%	45.5%	1306	2104	2078	510	765	816
RGF [31]	96.4%	85.3%	87.4%	1146	2088	2087	667	1280	1305
P-RGF [8]	94.5%	83.9%	85.9%	883	1583	1581	448	**657**	**690**
Meta Attack [12]	71.1%	33.8%	36%	3789	4101	4012	3202	3712	3649
Bandits [20]	99.2%	94.1%	95.3%	964	1737	1662	520	954	1014
Simulator Attack	99.4%	96.8%	97.9%	**811**	**1380**	**1445**	**431**	850	878

Untargeted attack under ℓ_∞ norm attack in TinyImageNet dataset

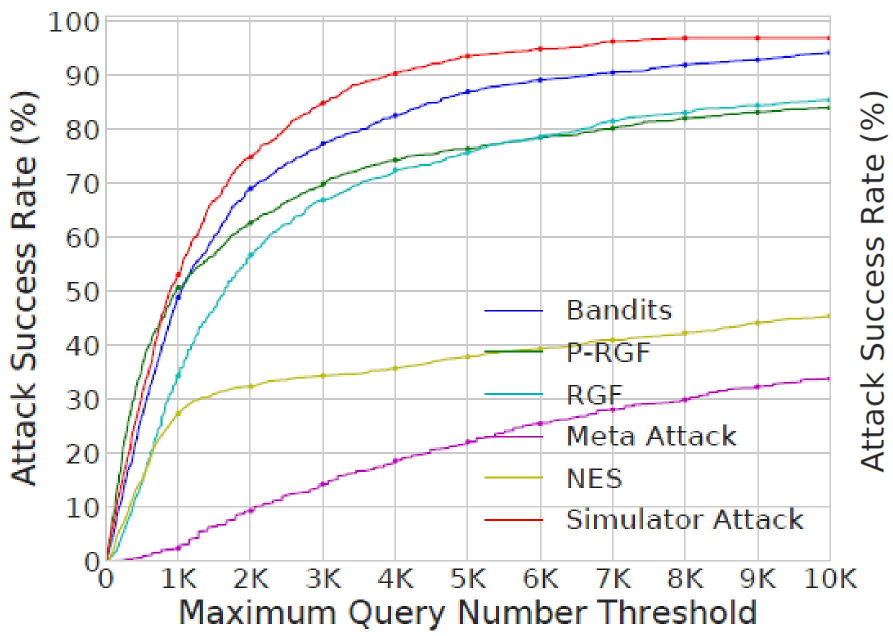
Attack	Attack Success Rate			Avg. Query			Median Query		
	D_{121}	R_{32}	R_{64}	D_{121}	R_{32}	R_{64}	D_{121}	R_{32}	R_{64}
NES [19]	88.5%	88%	88.2%	4625	4959	4758	4337	4703	4440
Meta Attack [12]	24.2%	21%	18.2%	5420	5440	5661	5506	5249	5250
Bandits [20]	85.1%	72.2%	72.4%	2724	3550	3542	1860	2700	2854
Simulator Attack	89.8%	84.9%	83.9%	**1959**	**2558**	**2488**	**1399**	**1966**	**1982**

Targeted attack under ℓ_2 norm attack in TinyImageNet dataset

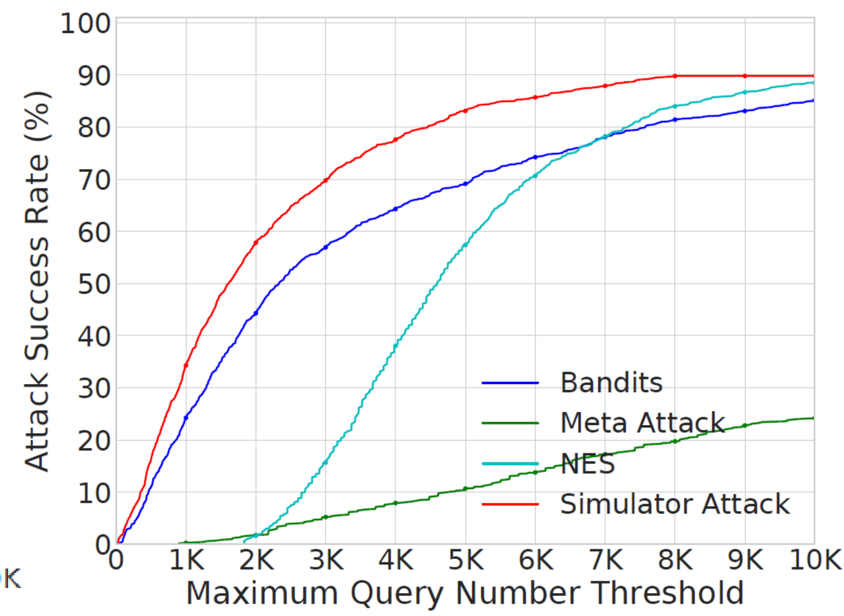
Comparisons with SOTA Methods under different maximum queries



PyramidNet-272 in CIFAR-100



ResNext-101 (32×4d) in TinyImageNet



DenseNet-121 in TinyImageNet

Results of attacks on defensive models

Dataset	Attack	Attack Success Rate				Avg. Query				Median Query			
		CD [21]	PCL [30]	FD [25]	Adv Train [28]	CD [21]	PCL [30]	FD [25]	Adv Train [28]	CD [21]	PCL [30]	FD [25]	Adv Train [28]
CIFAR-10	NES [19]	60.4%	65%	54.5%	16.8%	1130	728	1474	**858**	400	150	450	**200**
	RGF [31]	48.7%	82.6%	44.4%	22.4%	2035	1107	1717	973	1071	306	768	510
	P-RGF [8]	62.8%	80.4%	65.8%	22.4%	1977	1006	1979	1158	1038	230	703	602
	Meta Attack [12]	26.8%	77.7%	38.4%	18.4%	2468	1756	2662	1894	1302	1042	1824	1561
	Bandits [20]	44.7%	84%	55.2%	34.8%	786	776	832	1941	100	126	114	759
	Simulator Attack	54.9%	78.2%	60.8%	32.3%	**433**	**641**	**391**	1529	**46**	**116**	**50**	589
CIFAR-100	NES [19]	78.1%	87.9%	77.6%	23.1%	892	429	1071	**865**	300	150	250	**250**
	RGF [31]	50.2%	95.5%	62%	29.2%	1753	645	1208	1009	765	204	408	510
	P-RGF [8]	54.2%	96.1%	73.4%	28.8%	1842	679	1169	1034	815	182	262	540
	Meta Attack [12]	20.8%	93%	59%	27%	2084	1122	2165	1863	781	651	1043	1562
	Bandits [20]	54.1%	97%	72.5%	44.9%	786	321	584	1609	56	34	32	484
	Simulator Attack	72.9%	93.1%	80.7%	35.6%	**330**	**233**	**250**	1318	**30**	**22**	**24**	442
TinyImageNet	NES [19]	69.5%	73.1%	33.3%	23.7%	1775	863	2908	**945**	850	250	1600	**200**
	RGF [31]	31.3%	91.8%	9.1%	34.7%	2446	1022	1619	1325	1377	408	765	612
	P-RGF [8]	37.3%	91.8%	25.9%	34.4%	1946	1065	2231	1287	891	436	985	602
	Meta Attack [12]	4.5%	75.8%	3.7%	20.1%	1877	2585	4187	3413	912	1792	2602	2945
	Bandits [20]	39.6%	95.8%	12.5%	49%	893	909	1272	1855	85	206	193	810
	Simulator Attack	43%	84.2%	21.3%	42.5%	**377**	**586**	**746**	1631	**32**	**148**	**157**	632

Untargeted attack under ℓ_∞ norm attack in TinyImageNet dataset

CD: ComDefend

PCL: prototype conformity loss

FD: Feature Distillation

Conclusions

- **A novel black-box attack**
  - Improving the query efficiency by training a generalized substitute model.

- **A new type of security threat upon eliminating the target model in training.**
  - The adversary with the minimal information about the target model can also counterfeit this model.

- **A new way to use meta-learning**
  - The mean square error (MSE)-based knowledge-distillation loss carries out the inner and outer loops of meta-learning.
  - A query-sequence level partition strategy is adopted to divide each task into meta-train and meta-test sets.