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

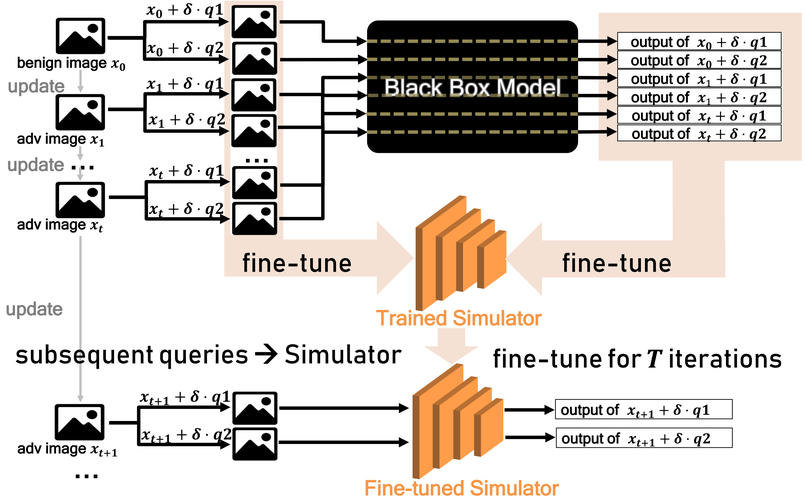
Chen Ma, Li Chen, and Jun-Hai Yong
School of Software, BNRist, Tsinghua University, Beijing, China

CVPR VIRTUAL JUNE 19-25

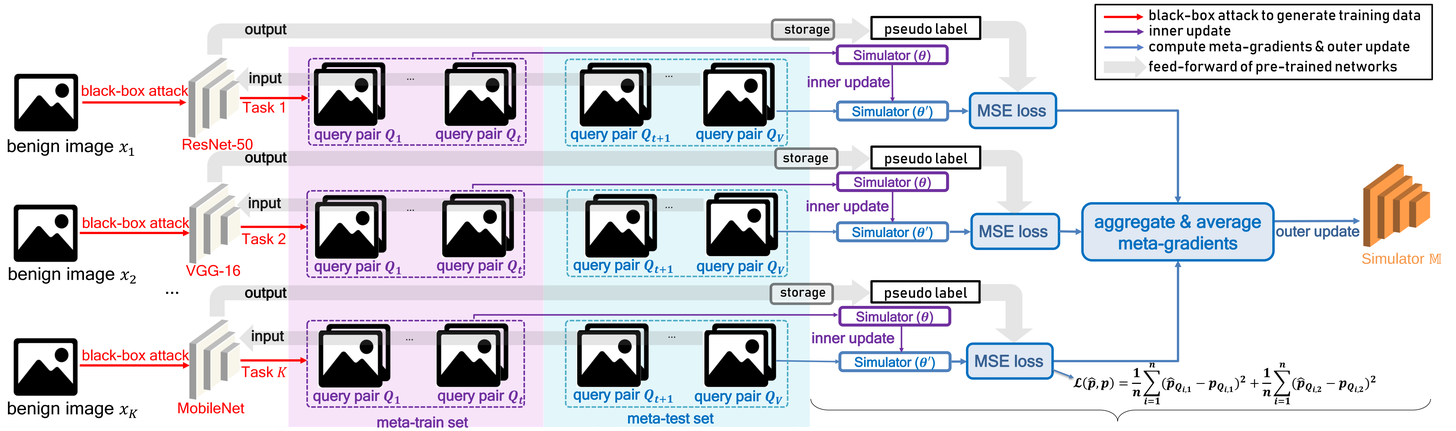
Introduction

The goal of this paper is to reduce the high query complexity of black-box attack by using a simulator. The contributions of this work are:

- We focus on training a generalized substitute model (named Simulator) without the target model requirement.
- The queries of the black-box attack are used as training data, thus allowing the Simulator to learn how to distinguish the subtle differences among queries.
- The training uses a knowledge-distillation loss to carry out the meta-learning between the Simulator and the many different existing classification networks.
- Once trained and fine-tuned, the Simulator can mimic the output of any target model that is unseen in training, enabling it to eventually replace the target model to transfer the query stress in the attack.



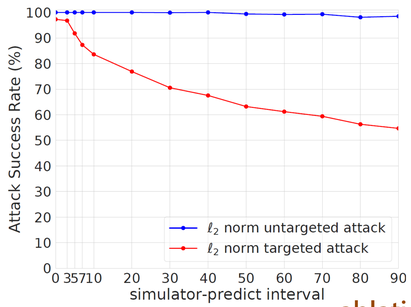
Method

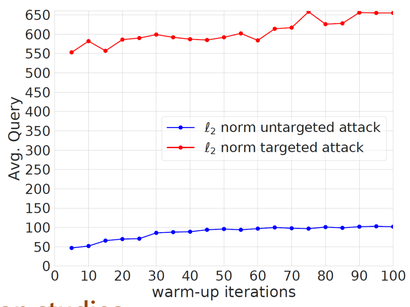


Query sequences generated during the attack (cached in the hard drive)

Simulator Learning

Experimental Results





ablation studies

meta-predict interval: a difficult attack (targeted attack) requires a small simulator-predict interval.

warm-up iterations: more warm-up iterations cause higher average queries.

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

ℓ_∞ norm attack on four defensive models: ComDefend (CD), Feature Distillation (FD), prototype conformity loss (PCL) and Adv Train. Simulator Attack performs the best among all methods.