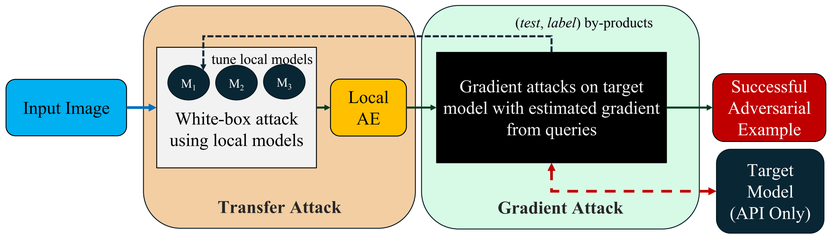
Hybrid Batch Attacks: Finding Black-box Adversarial Examples with Limited Queries

Fnu **Suya**, Jianfeng Chi, David Evans, Yuan Tian

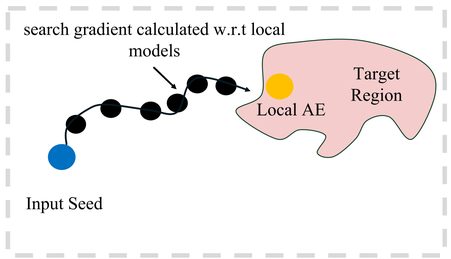
USENIX Security Symposium 2020

Goal: Estimate the cost for a black-box adversary to find adversarial examples.

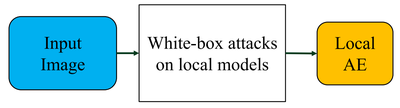
Combining Transfer and Gradient Attacks



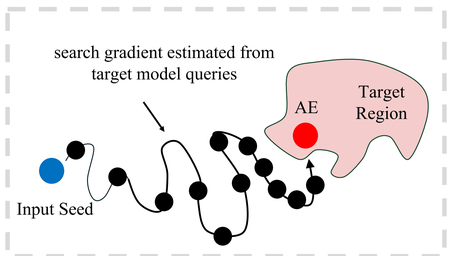
Search Space of Transfer Attack



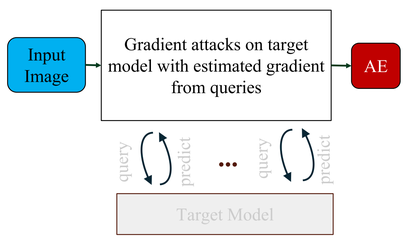
Process of Transfer Attack



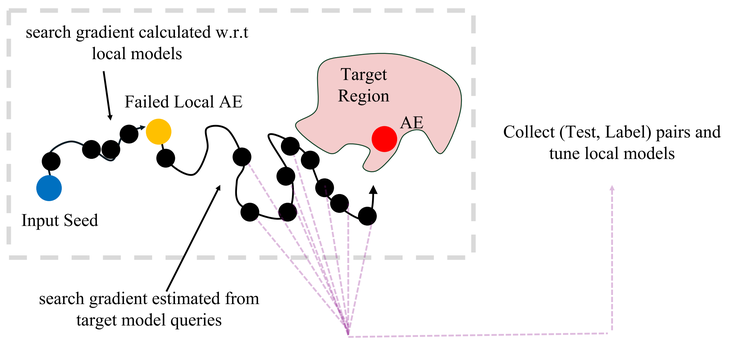
Search Space of Gradient Attack



Process of Gradient Attack



Search Space of Hybrid Attack



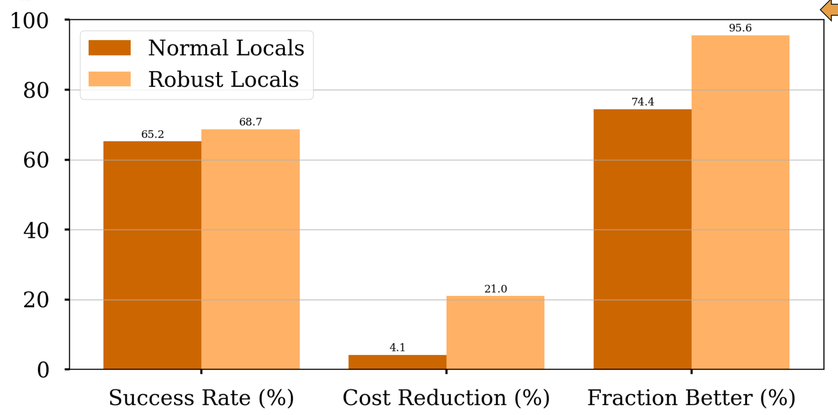
Collect (Test, Label) pairs and tune local models

Does it help gradient attacks to start from failed transfer candidates?

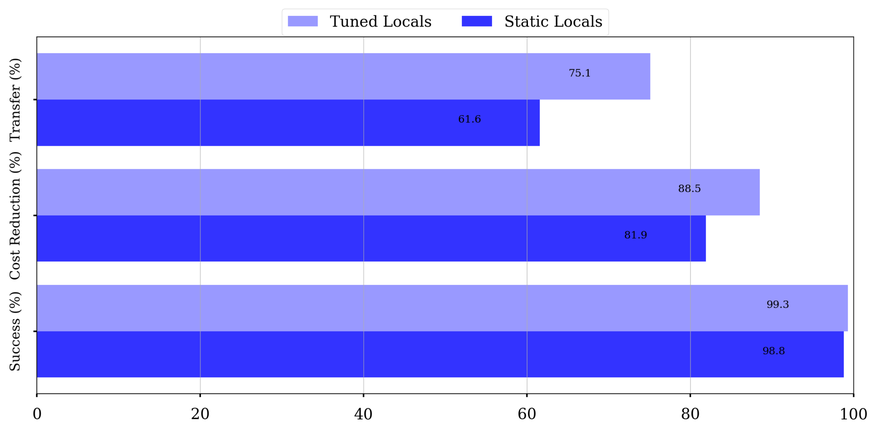
Model	Success Rate (%)		Queries/AE		Fraction Better (%)
	Base	Ours	Base	Ours	
MNIST Normal (Targeted)	90.9	**98.8**	1,645	**298**	99.8
CIFAR10 Normal (Targeted)	92.2	**98.1**	1,227	**277**	98.7
ImageNet Normal (Targeted)	93.6	**97.2**	42,417	**24,104**	91.8
CIFAR10 Robust (Untargeted)	64.4	**65.2**	2,640	**2,529**	74.4

AutoZOOM [AAAI 2018] gradient attack, local models: normal models

Impact of local models:



Using Byproducts to Tune Local Models



To attack MNIST models, tuning the local models helps to improve the attack performance. However, for CIFAR10 models, we observe degradation of attack performance by tuning the local models.

Batch Attack

Can we do better when total query is limited?

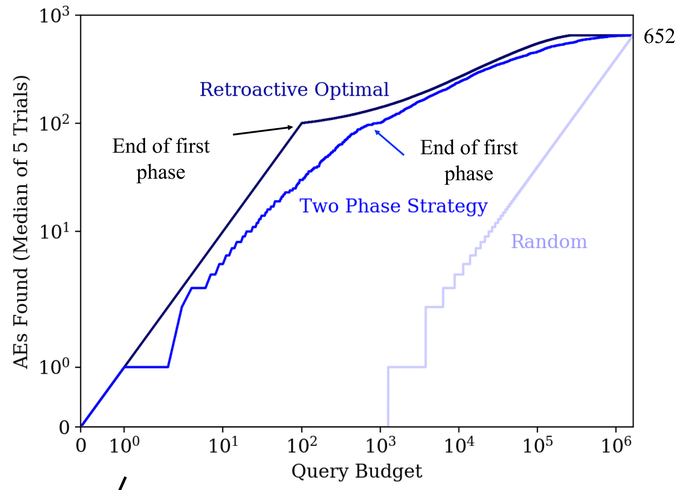
Hybrid Batch Attack

Goal: prioritize low-cost-to-attack seeds
Two Phase Strategy:

***First phase*:** find *direct transfers* by sorting images based on *local PGD-Steps*

***Second phase*:** find *easy images* by sorting based on *target model attack loss*

Evaluating Hybrid Batch Attack



AutoZOOM gradient attack on robust CIFAR10 model, local models are normal models

Number of queries to obtain **10%** of 1000 seeds:
Optimal: 100; **Two-Phase**: 824; **Random**: 251,682

NSF Center for Trustworthy Machine Learning
https://ctml.psu.edu/