Hybrid Batch Attacks

Finding Black-box Adversarial Examples with Limited Queries

Fnu Suya
Jianfeng Chi
David Evans
Yuan Tian

University of Virginia

USENIX Security 2020

evadeML.org

Two Types of Black-box Attacks

	Query Cost	Success Rate
Optimization Attacks	High	**High**
Transfer Attacks	**Low**	Low

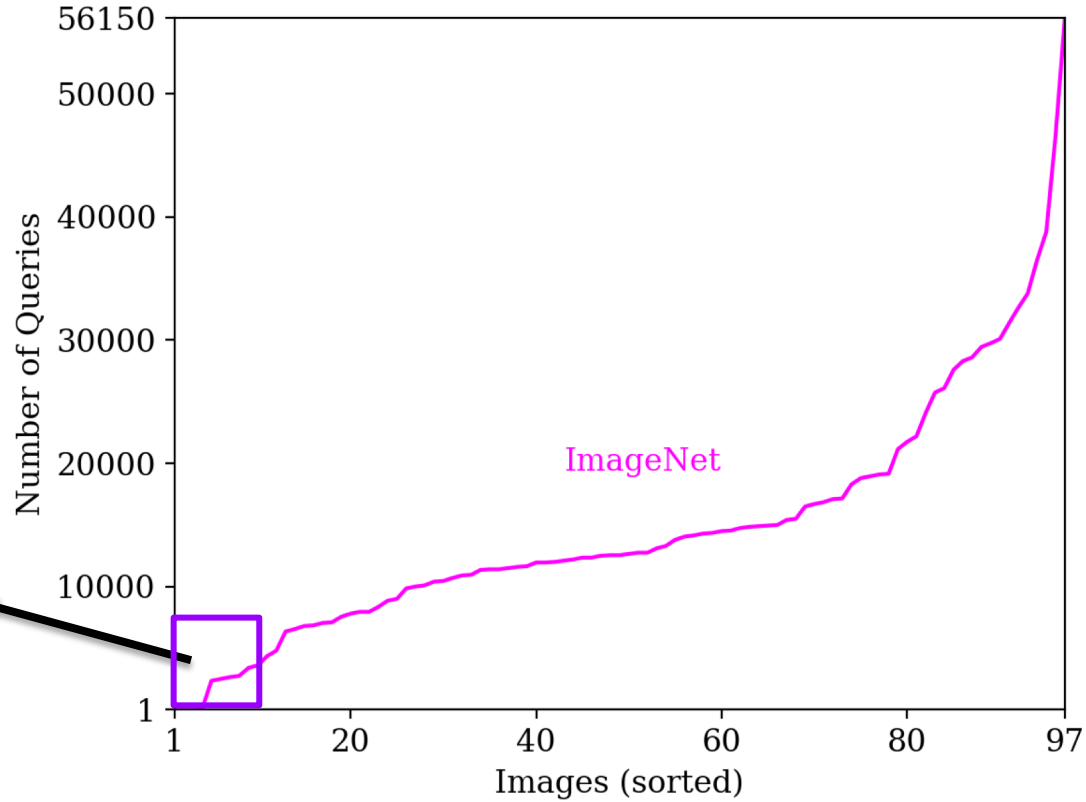
Can we combine the attack strategies to get high success and low cost?

Result Summary: Hybrid Attack

Model	Attack Success Rate (%)		Query Cost	
	Baseline Optimization Attack	Hybrid Attack	Baseline Optimization Attack	Hybrid Attack
Standard CIFAR10	92.2	**98.1**	1,227	**277**
Robust CIFAR10	64.3	**68.7**	2,640	**2,068**
Standard ImageNet	93.6	**97.2**	42,417	**24,104**

Attacks Evaluated by *Average* Cost

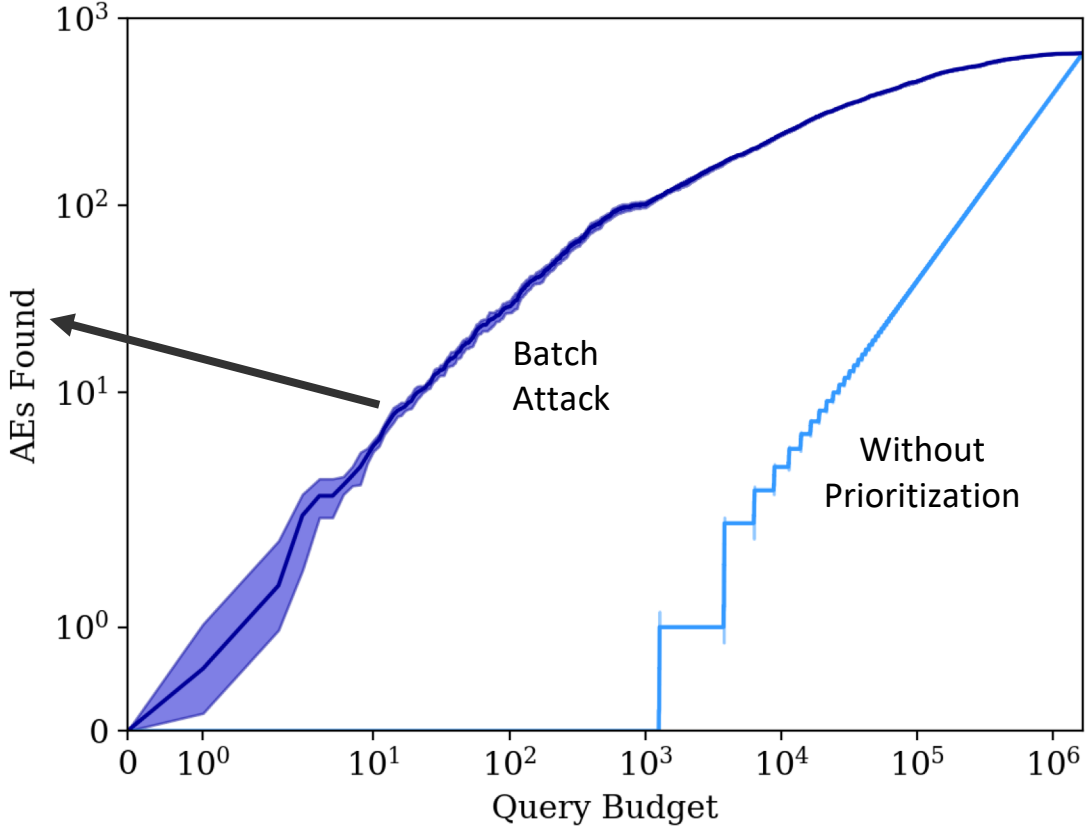
Top 10 mean: 3,402
Overall mean: 13,722
Potential Savings: 75%



Result Highlight: Batch Attack

Batch Attack	Without Prioritization
20	**24,054**

Total queries to find 10 adversarial examples (given 1000 candidate seeds)



Rest of Talk

Hybrid Attacks

How to combine transfer and optimization attacks?

Attackers can exploit all known attacks

Batch Attacks

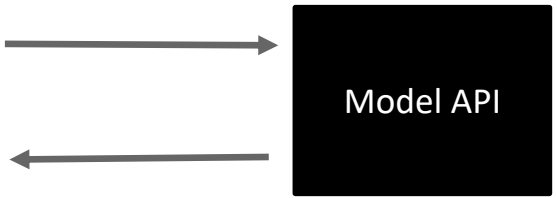
How to efficiently find low-cost seeds?

Attackers can choose best candidate seeds to attack

Relax assumptions to better estimate attack cost for realistic adversaries

Our Threat Model

Black-box: only **query access** to model without internal information



“pig”: 0.84
“dog”: 0.02
…

Queries are expensive (financial cost or detection risk)

Google Vision API: first 1000 queries (free), $1.5/1000 queries

Amazon Face Recognition: $1.0/1000 queries

Clarifai (NSFW): first 5000 queries (free), $3.2/1000 queries for custom model

Attacks in our experiments would average **$42-$120 per adversarial example found**

Transfer Attacks

Goodfellow et al. (2014)

Papernot et al. (2017)

Liu et al. (2017)

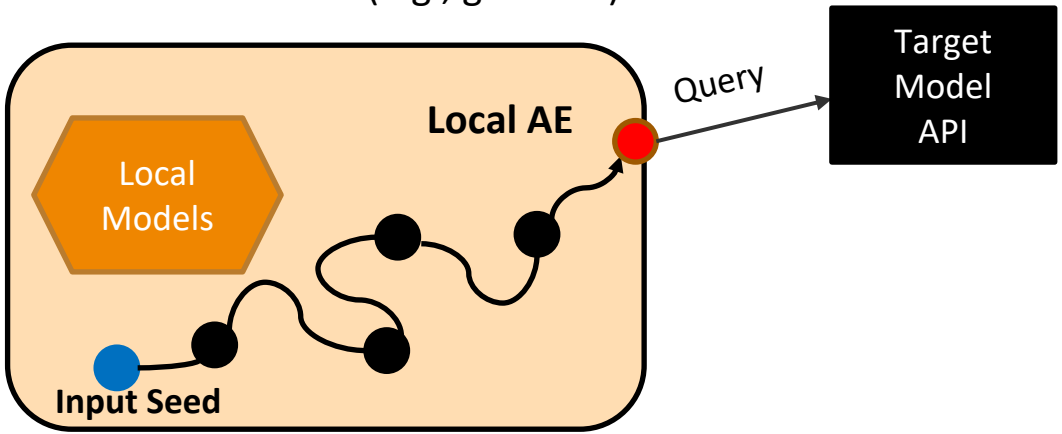
Dong et al. (2018)

Xie et al. (2019)

...

Li et al. (2020)

Perturbation is generated with local model information (e.g., gradient)



Attack Search Space

Low success (transfer) rate for harder attack settings

Optimization Attacks

Zhang et al. (2017)

Ilyas et al. (2018)

Bhagoji et. al (2019)

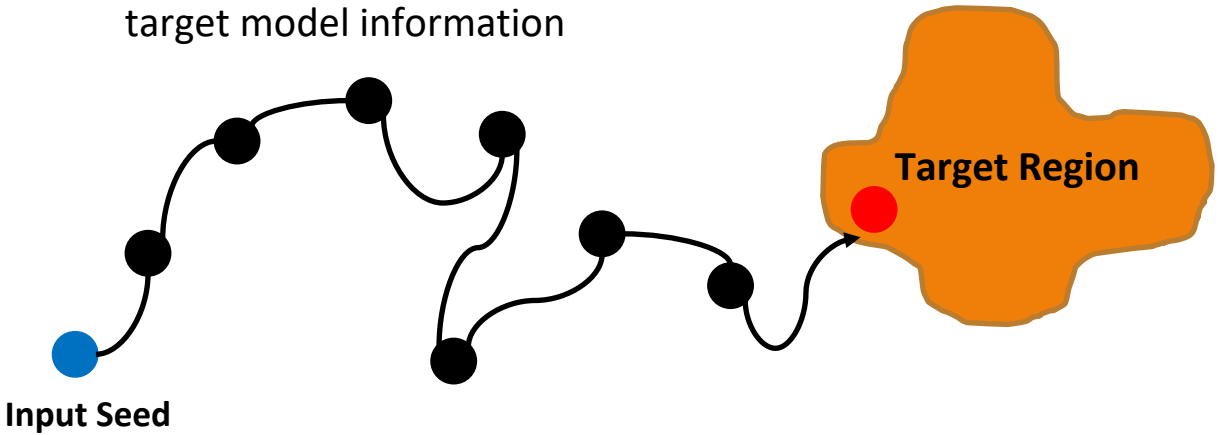
Tu et al. (2019)

Moon et al. (2019)

...

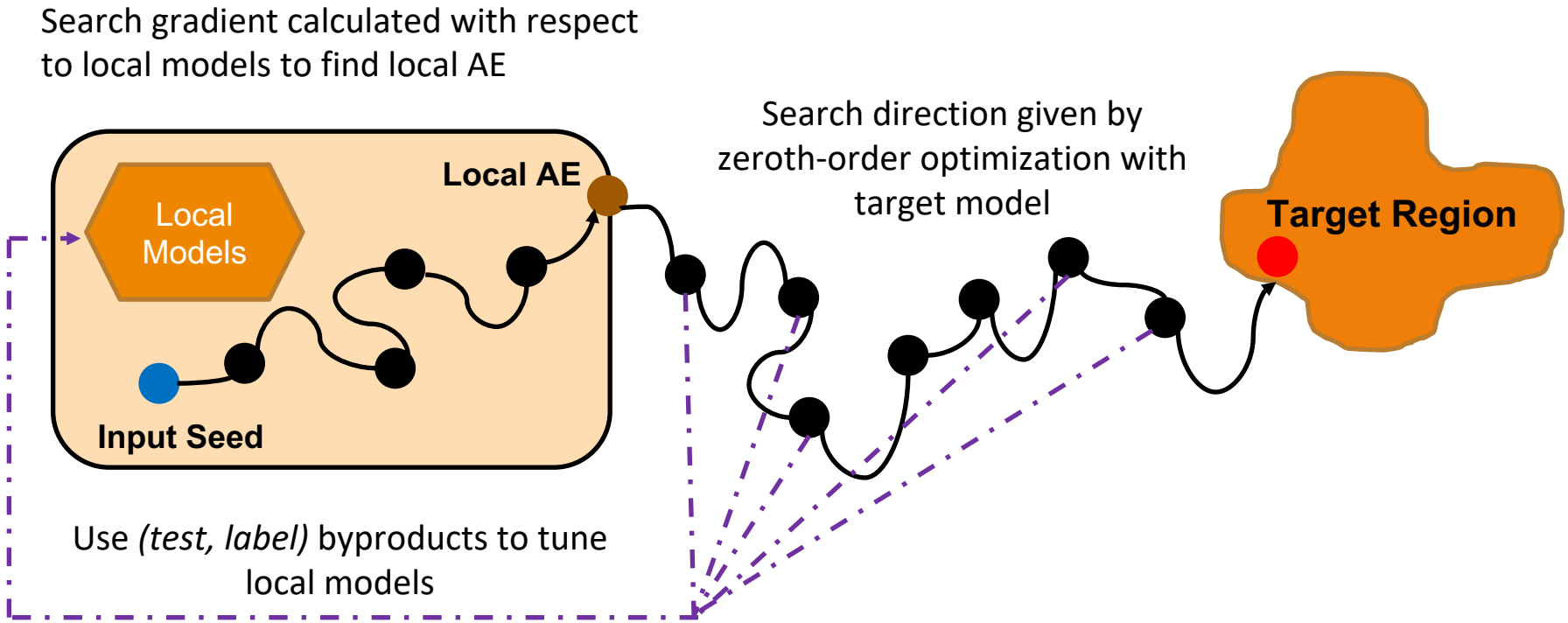
Andriushchenko et al. (2020)

Perturbation is given by zeroth-order optimization with target model information



Attack Search Space

Combine Transfer and Optimization Attacks



(See paper for byproduct direction, which is usually not successful)

Local Adversarial Example Generally Helps for Standard Target Models

Standard Target Model	Success Rate (%)		Query Cost		Fraction Better (%)
	Optimization	Hybrid	Optimization	Hybrid	
MNIST [1]	90.9	**98.8**	1,645	**298**	**99.8**
CIFAR10 [1]	92.2	**98.1**	1,227	**277**	**98.7**
ImageNet [1]	93.6	**97.2**	42,417	**24,104**	**91.8**
ImageNet [2]	73.0	**98.0**	31,849	**6,840**	**100.0**

Optimization Attack: AutoZOOM [1], SimBA [2]; **Transfer Attack**: PGD on ensemble [3]; **Local Models**: Standard; **Target Class:** Least Likely Class

[1] Tu, Chun-Chen, et al. “Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks.” (AAAI 2019).

[2] Guo, Chuan, et al. "Simple black-box adversarial attacks." (ICML 2019).

[3] Liu, Yanpei, et al. "Delving into transferable adversarial examples and black-box attacks." (ICLR 2017).

Except Against Robust Target Model

Robust Target Model	Success Rate (%)		Query Cost		Fraction Better (%)
	Optimization	Hybrid	Optimization	Hybrid	
CIFAR10 [4] (Untargeted)	64.4	**65.2**	2,640	**2,529**	**74.4**

Optimization Attack: AutoZOOM; **Transfer Attack**: PGD on ensemble; **Local Models**: Standard;

[4] Madry, Aleksander, et al. “Towards deep learning models resistant to adversarial attacks.” (ICLR 2018).

Failure on Robust Target Model

Hypothesis: *different vulnerability space of standard and robust models*

	Transfer Rate (%)		Success Rate (%)		Cost Reduction (%)	
	Standard Local	Robust Local	Standard Local	Robust Local	Standard Local	Robust Local
Standard Target	**63.6**	18.4	**98.2**	95.3	**77.1**	35.7
Robust Target	10.1	**40.7**	65.3	**68.7**	3.8	**20.5**

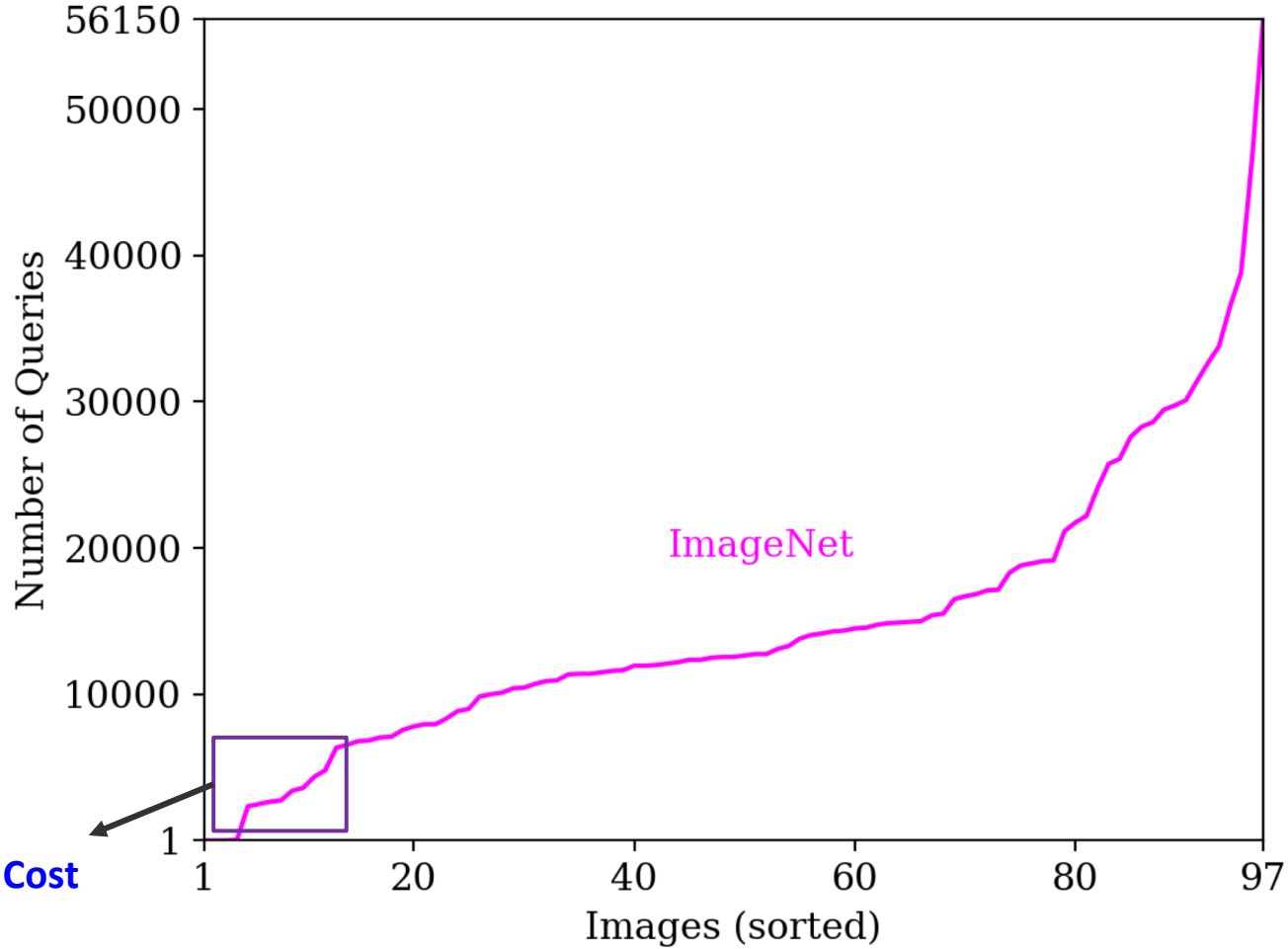
Local and target models of same type match best

Failed to find universal local models (see paper for details)

Main Takeaway from Hybrid Attack

Starting from local failed transfers reduces cost of optimization attacks

So far: reducing *average* cost

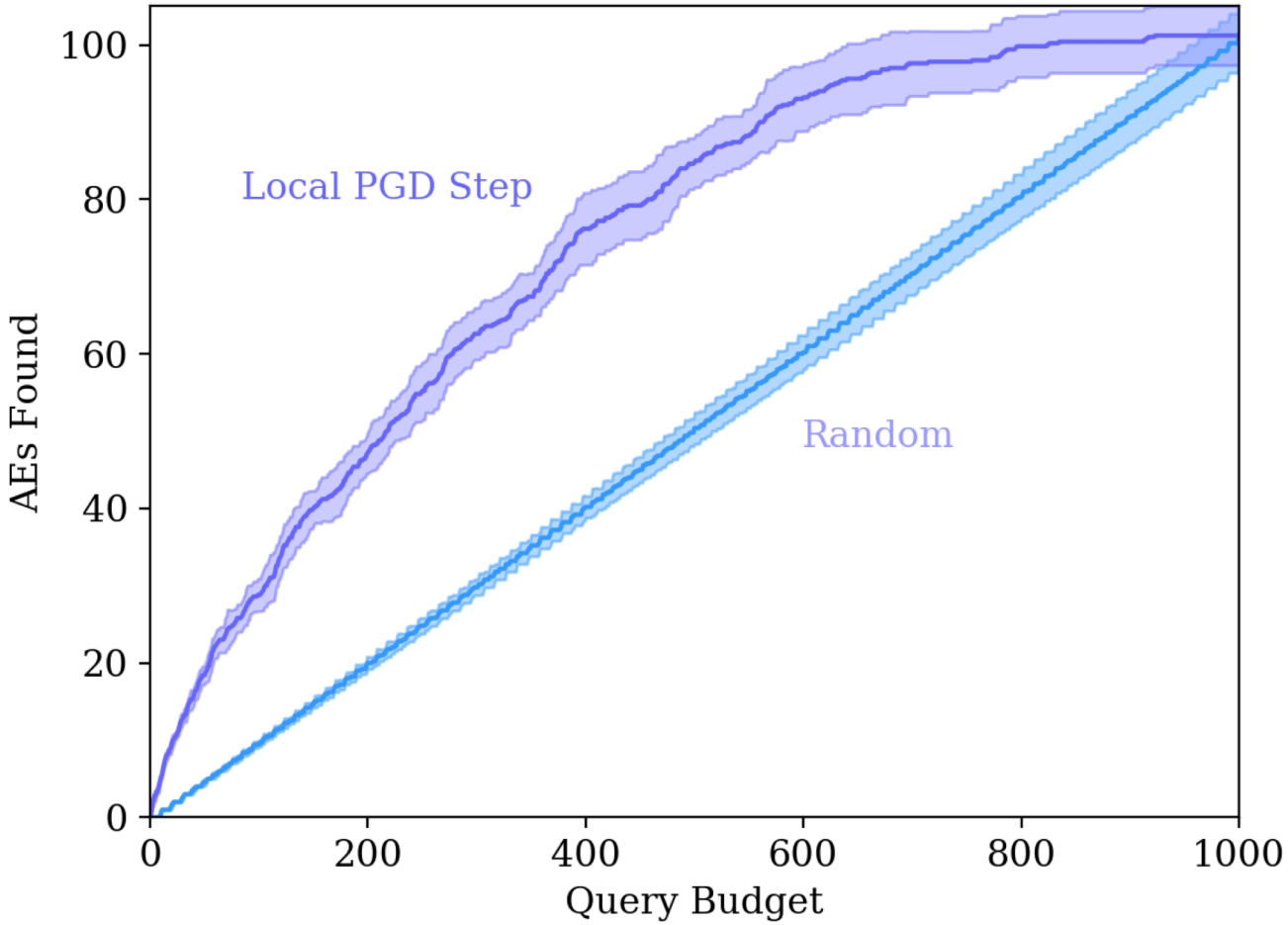


Low Query Cost

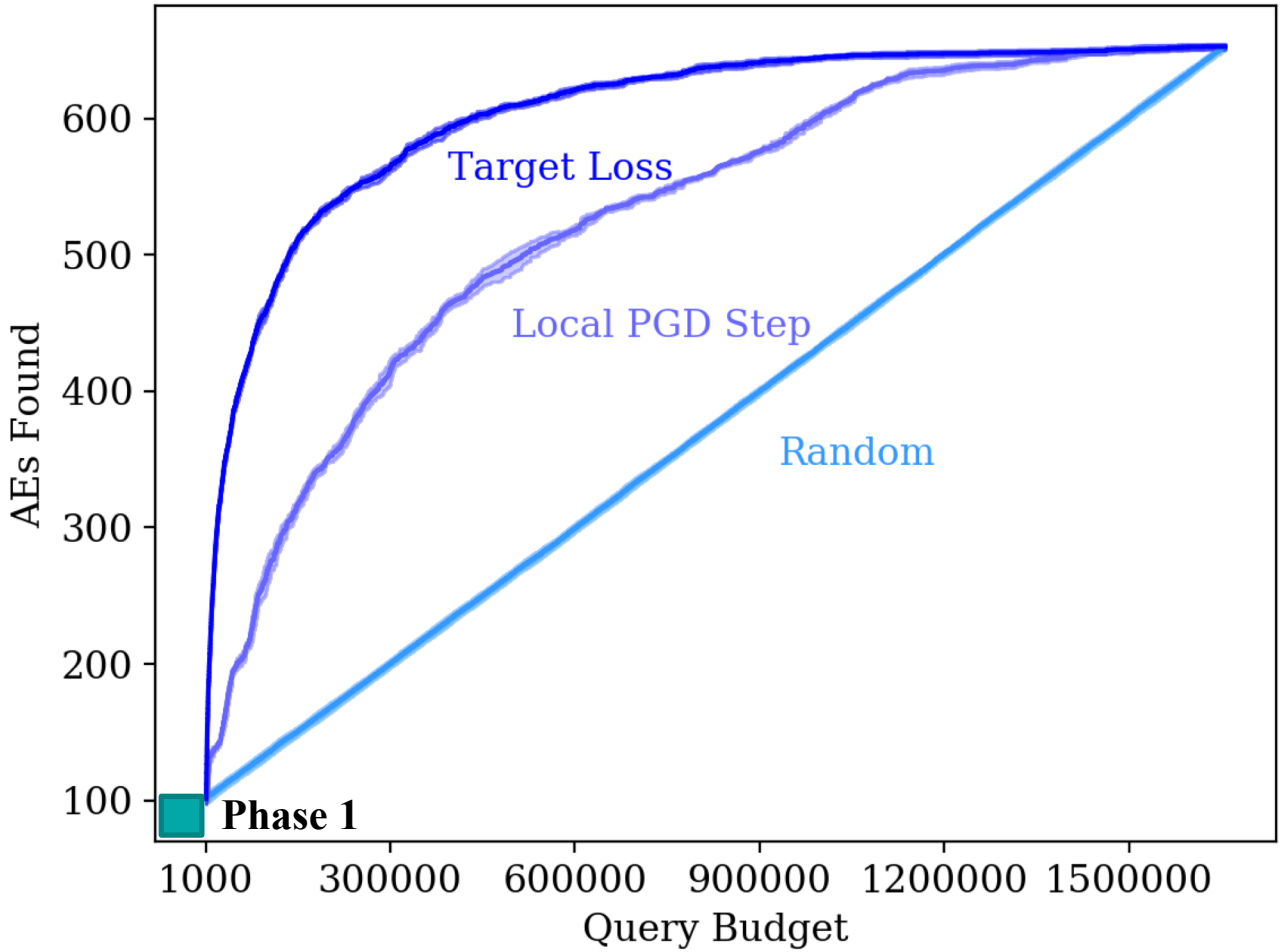
How can we efficiently find those *low-cost* seeds?

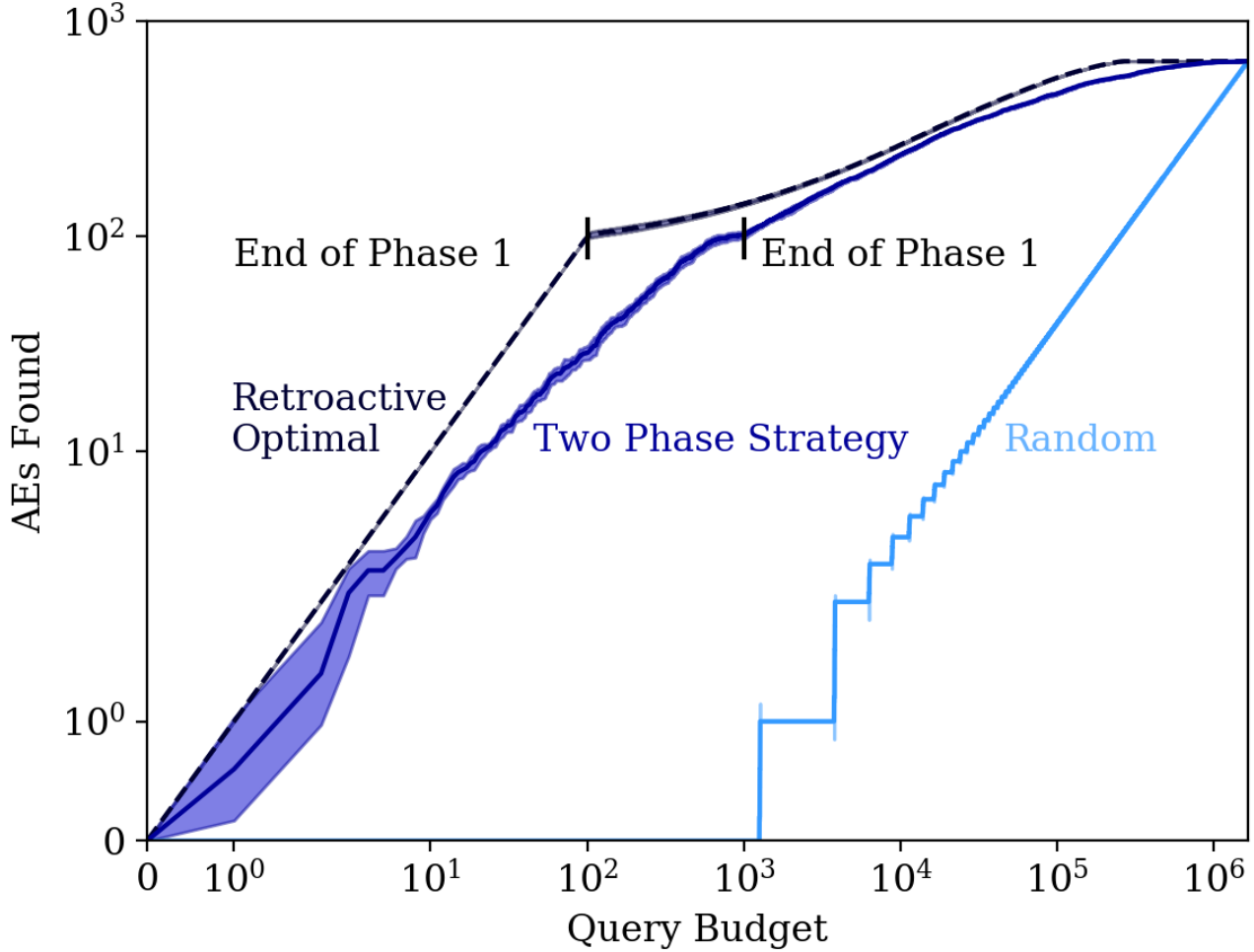
Phase 1: Transfer	Phase 2: Optimization
Information available: Results of attempted attack using local models	**Information available:** Results of previous attempts for transferability
Hypothesis: If local models find successful adversarial example easily, more likely to transfer.	**Hypothesis:** Lower loss values are closer to the target region and easier to attack.
Strategy: Prioritize seeds with fewer attack iterations	**Strategy:** Prioritize seeds with lower loss function values

Phase 1: Direct Transfers



Phase 2: Optimization Attack




10^3
10^2
10^1
10^0
0
AEs Found
End of Phase 1
End of Phase 1
Retroactive
Optimal
Two Phase Strategy
Random
0
10^0
10^1
10^2
10^3
10^4
10^5
10^6
Query Budget

Performance of Two-Phase Strategy

Prioritization Method	1%	5%	10%
Retroactive Optimal	10	50	108
Two-Phase	**20**	**218**	**826**
Random	24,054	125,327	251,917

Number of queries needed to get different fraction of 1000 images

Target: Robust CIFAR10, **Local**: Standard CIFAR10, **Attack**: AutoZOOM, Averaged 5 Times
(see paper for standard errors)

Open Source Implementation

https://github.com/suyeecav/Hybrid-Attack

Tutorials of incorporating new attacks

Supports both TensorFlow and PyTorch

Can be applied to decision-based attacks

Main Takeaway

Understanding *cost* of an attack, requires considering realistic adversaries who can pick and choose *what* and *how* to attack to achieve their goals

Hybrid Attack: combine attacks
Batch Attack: seek easy images

code: **https://github.com/suyeecav/Hybrid-Attack**

updated paper: **https://arxiv.org/abs/1908.07000**

contact: Fnu Suya **suya@virginia.edu**

Fnu Suya

Jianfeng Chi

David Evans

Yuan Tian

Contributors: Emily Buerk, Jessie Li, Konrad Siebor, Brian Tran

UNIVERSITY of VIRGINIA