NEURAL INFORMATION
PROCESSING SYSTEMS

sAIfer Lab
Joint lab on Safety and Security of AI

TransferBench: Benchmarking Ensemble-based Black-box Transfer Attacks

Fabio Brau, Maura Pintor, Antonio Emanuele Cinà, Raffaele Mura, Luca Scionis,
Luca Oneto, Fabio Roli, Battista Biggio

39th Conference on Neural Information Processing Systems, San Diego, CA, USA

UNICA
UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

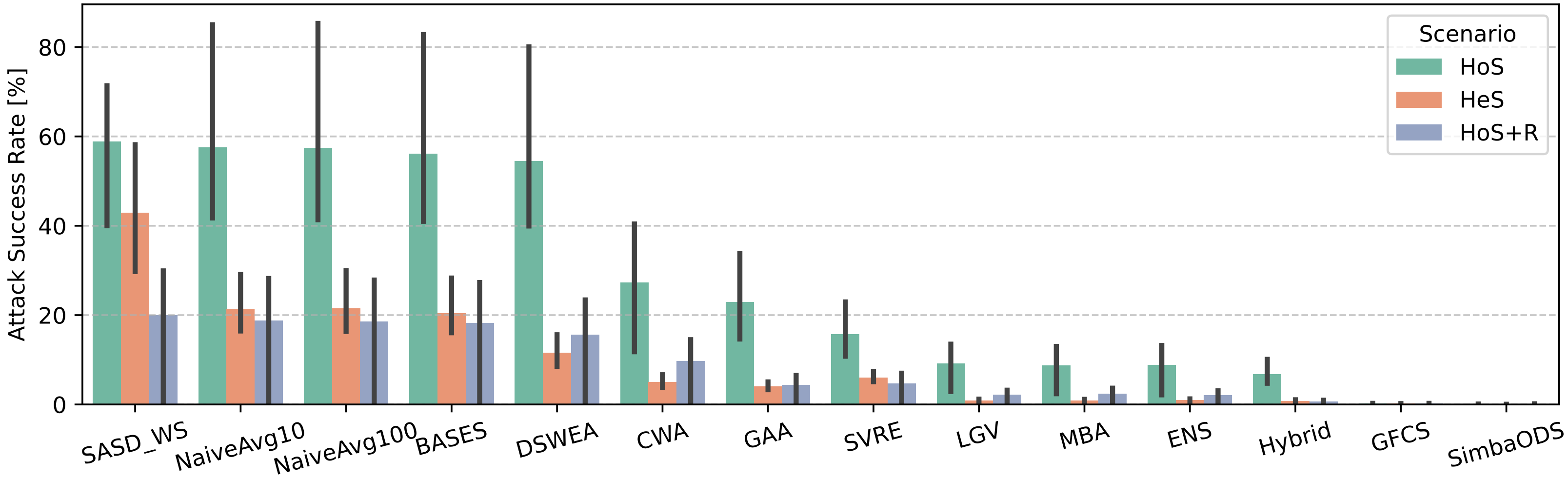
Università
di Genova

SAPIENZA
UNIVERSITÀ DI ROMA

Benchmarking Attacks on Standard Scenarios

TransferBench

Benchmarking Ensemble-based Black-box
Transfer Attacks

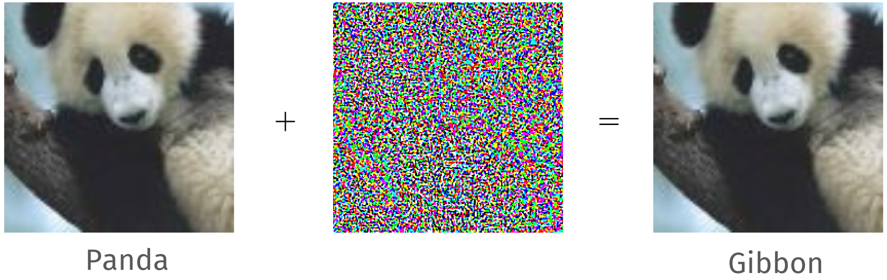


Attacking Classification Models

Gradient-Based Perturbation

$$x^* = x - \varepsilon\,\mathrm{sgn}\,\boxed{\nabla_x \mathcal{L}(g_\theta(x), y)}$$

Gradient based on the target



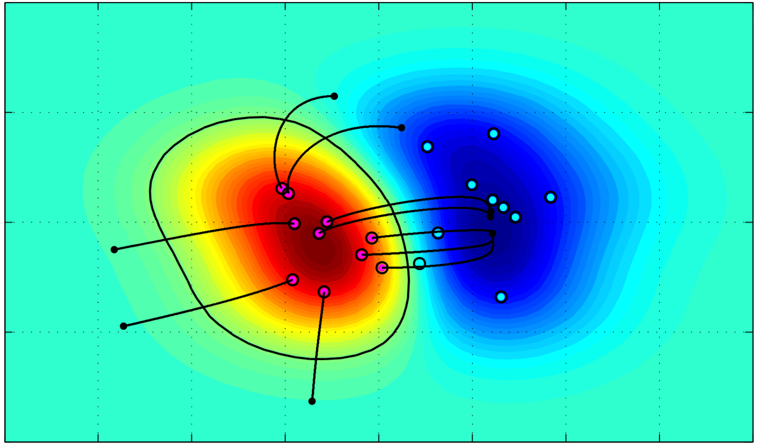
Goodfellow, Ian J., et el. "Explaining and harnessing adversarial examples.”. ICLR, 2015.

Attacking Classification Models

Adversarial Attacks as Minimum Problem

$$
\begin{aligned}
x^* \in \operatorname*{argmin}_{x} \quad & \boxed{\mathcal{L}(g_\theta(x), y)} \\
\text{s.t.} \quad & d(x, x_0) \leq \varepsilon
\end{aligned}
$$

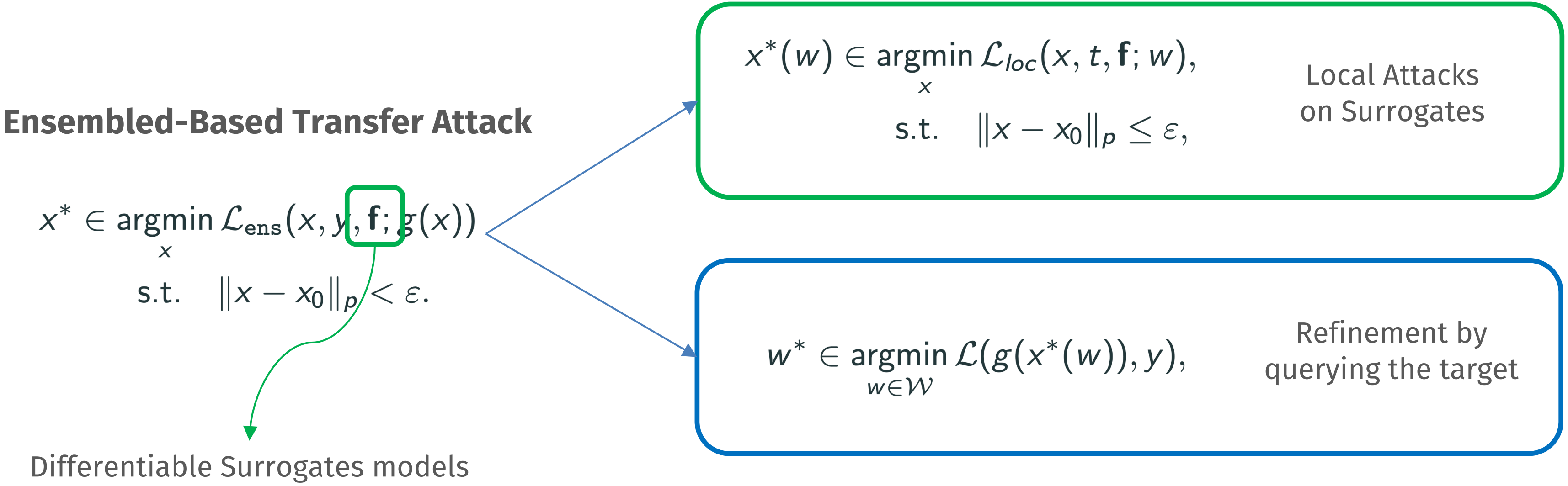
Assuming Differentiable Objective



Biggio, Battista, et al. "Evasion attacks against machine learning at test time." ECML-PKDD, 2013.

Ensemble-based Attacks Formulation

With a Black-box Target, gradient is not accessible ~~$\nabla_x \mathcal{L}(g_\theta(x), y)$~~



Ensembled-Based Transfer Attack

$$x^* \in \underset{x}{\operatorname{argmin}}\, \mathcal{L}_{\mathtt{ens}}(x, y \boxed{, \mathbf{f};} g(x))$$
$$\text{s.t.} \quad \|x - x_0\|_p < \varepsilon.$$

Differentiable Surrogates models

$$x^*(w) \in \underset{x}{\operatorname{argmin}}\, \mathcal{L}_{loc}(x, t, \mathbf{f}; w),$$
$$\text{s.t.} \quad \|x - x_0\|_p \leq \varepsilon,$$

Local Attacks on Surrogates

$$w^* \in \underset{w \in \mathcal{W}}{\operatorname{argmin}}\, \mathcal{L}(g(x^*(w)), y),$$

Refinement by querying the target

Coverage of the Benchmark and Motivation

Which is the best Ensemble-Based Transfer Attack ?

Compared Methods

Attack	Venue	m
SubSpace	NeurIPS 2019	3
SimbaODS	NeurIPS 2020	4
Hybrid	Usenix 2020	3
GFCS	ICLR 2022	4
BASES	BASES 2022	20
GAA	PR 2024	4
DSA	Usenix 2024	3
DSWEA	PR 2025	10

Large pool of surrogates has been sometime used !!

Flaws of Current Method

- Biased Surrogates
- Weak Targets
- Query Effectiveness

How TransferBench Addresses the Gaps

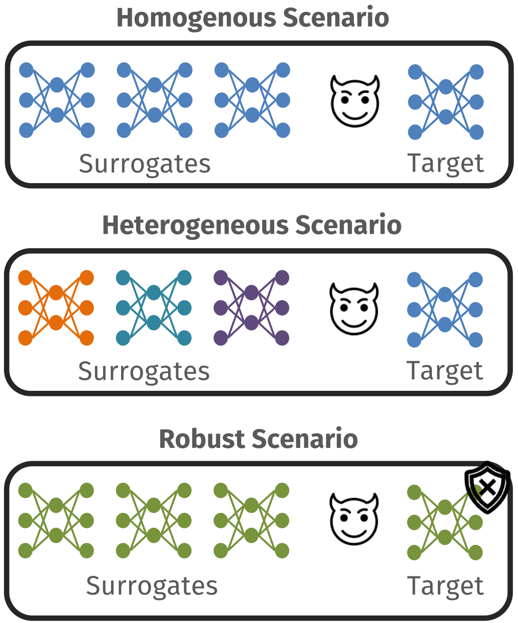
Flaws

Mitigations

Biased Surrogates

Weak Targets

Scenarios



Homogenous Scenario

Heterogeneous Scenario

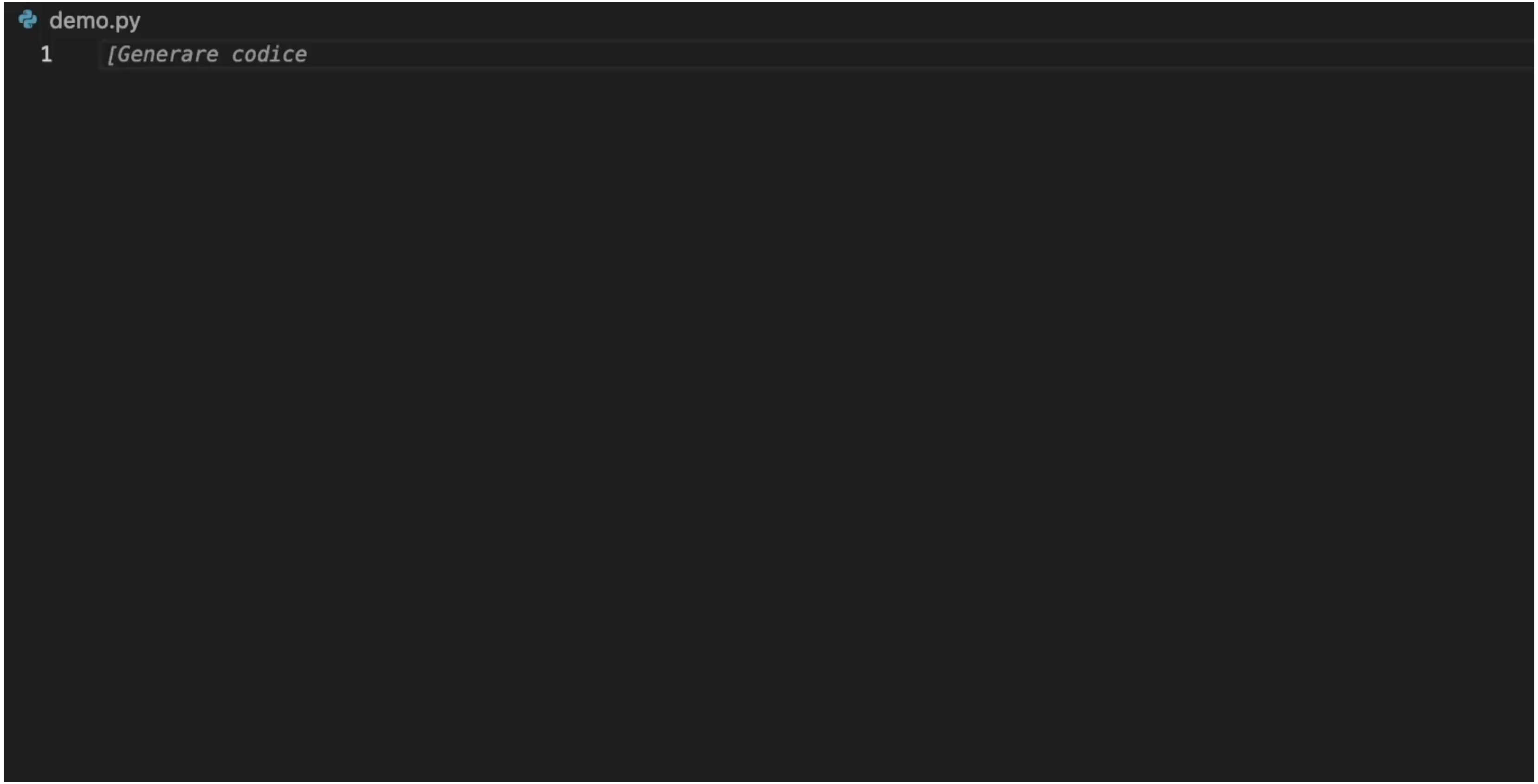
Robust Scenario

Query Effectiveness

Baselines

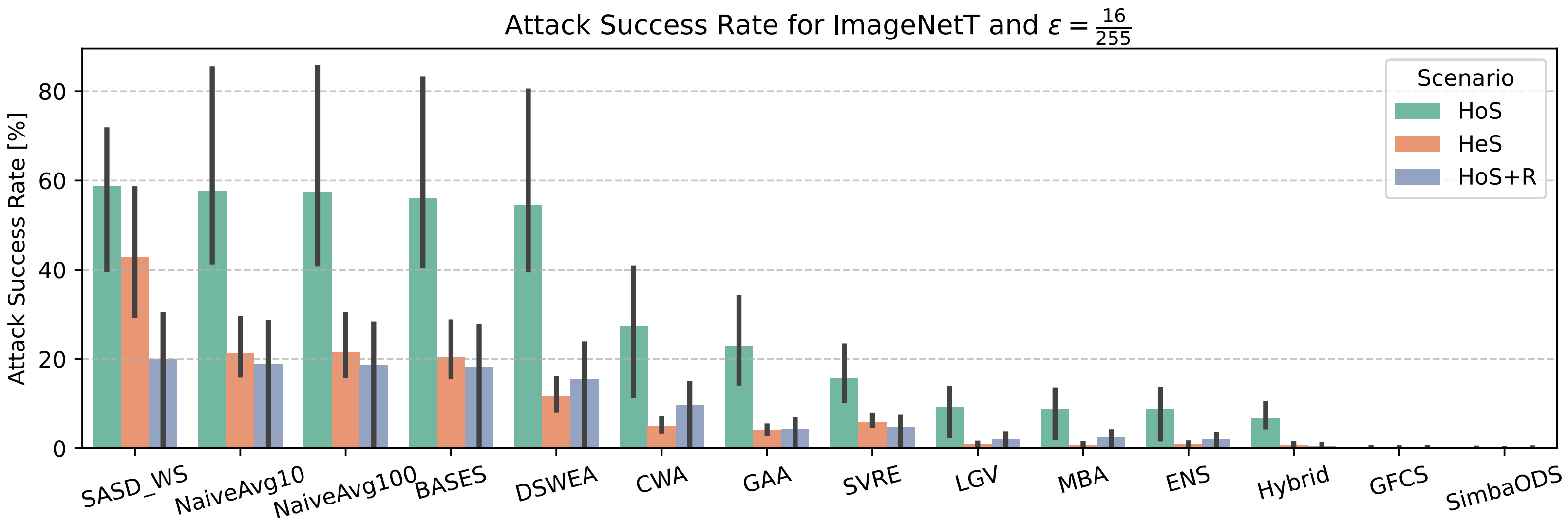
Query-free Methods and Naïve Average

Transferbench Ease of Use



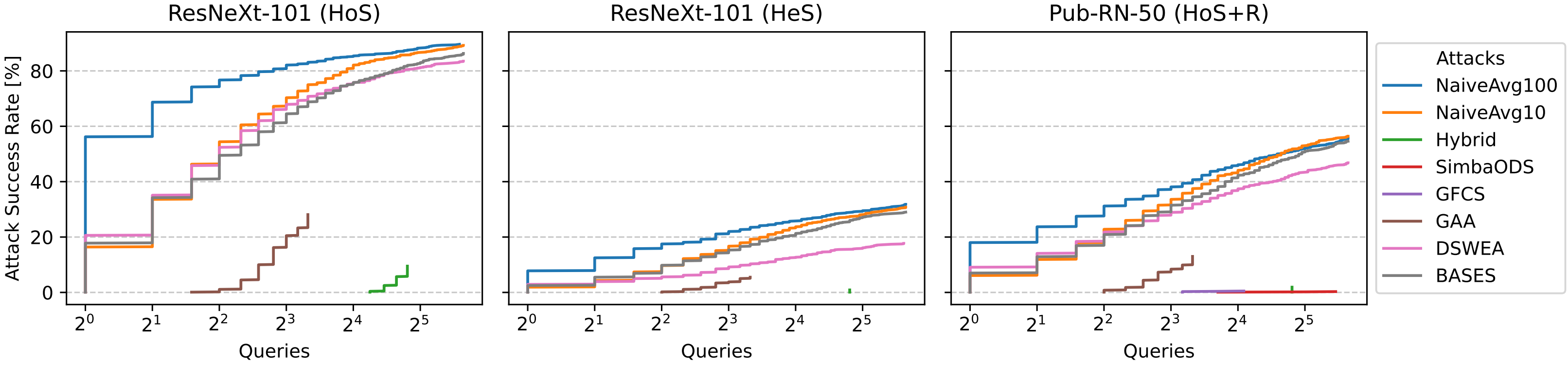
```
[Generare codice
```

Main Results



- In all the Scenarios, current methods are worst than baselines
- Attacking Robust Models is still an open problem

Main Results



Querying the target does not really contribute to refine the attack

NEURAL INFORMATION PROCESSING SYSTEMS

TransferBench

Benchmarking Ensemble-based Black-box Transfer Attacks

Paper

Code

SEC4AI4SEC

FAIR Future Artificial Intelligence Research

SERICS SECURITY AND RIGHTS IN THE CYBERSPACE

elsa European Lighthouse on Secure and Safe AI