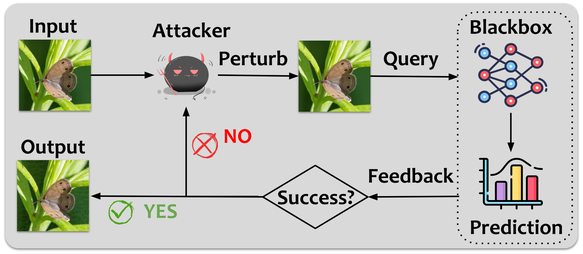
Blackbox Attacks via Surrogate Ensemble Search

Zikui Cai, Chengyu Song, Srikanth V. Krishnamurthy, Amit K. Roy-Chowdhury, M. Salman Asif
University of California, Riverside

Introduction

Blackbox attacks is a practical setting where the attacker does not have access to the internal parameters of the victim model.



Existing blackbox attack approaches can be generally categorized into the following categories.

Transfer-based attacks

Attack surrogate model(s) and test the perturbation on the victim models.

- ✓ Do not need feedback
- ✗ Low success rate due to model difference

Query-based attacks

Update perturbations by iteratively querying the victim model.

- ✓ High success rate
- ✗ High query count due to gradient estimation

Transfer-based combined with query

Query in a potentially low dimensional transferable search space.

- ✓ Better overall performance
- ✗ Targeted attack remains challenging (hundreds of queries)

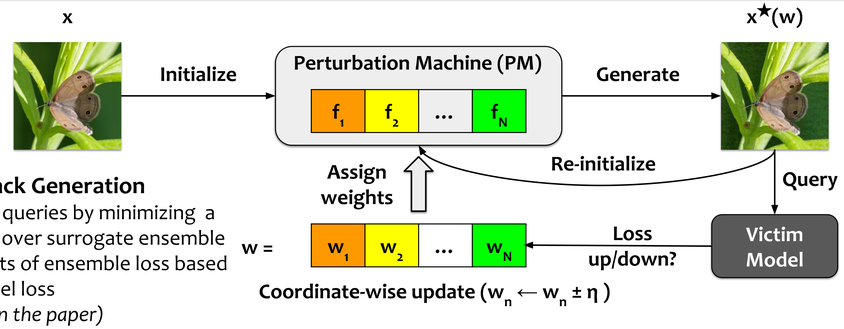
References

1. **(P-RGF)** Cheng et al. Improving black-box adversarial attacks with a transfer-based prior. NeurIPS 2019
2. **(TREMBA)** Huang et al. Black-box adversarial attack with transferable model-based embedding. ICLR 2019
3. **(ODS)** Tashiro et al. Diversity can be transferred: Output diversification for white-and black-box attacks. NeurIPS 2020
4. **(GFCS)** Lord et al. Attacking deep networks with surrogate-based adversarial black-box methods is easy. ICLR 2022

Blackbox Attacks via Surrogate Ensemble Search (BASES)

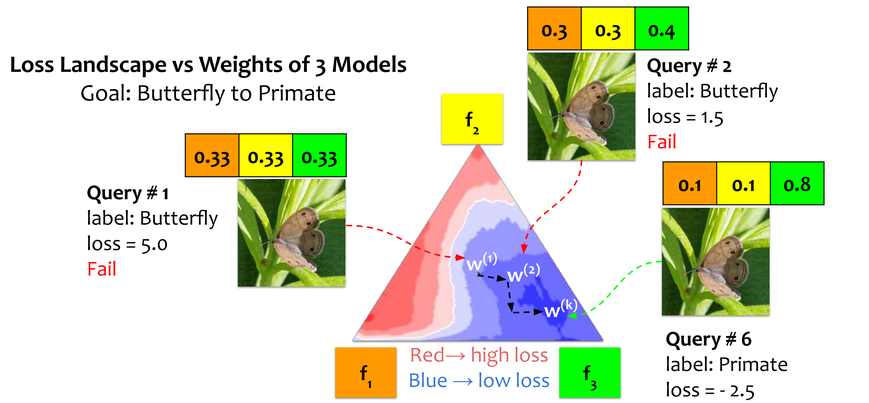
Generates attacks for blackbox victim model by searching weights of surrogate ensemble

- ✓ **Highly successful attacks** (by adapting perturbation to victim model)
- ✓ **Extremely small number of queries** (search space is low-dimensional)



Attack Generation

1. PM generates queries by minimizing a weighted loss over surrogate ensemble
2. Update weights of ensemble loss based on victim model loss
   (More details in the paper)



Experimental Results

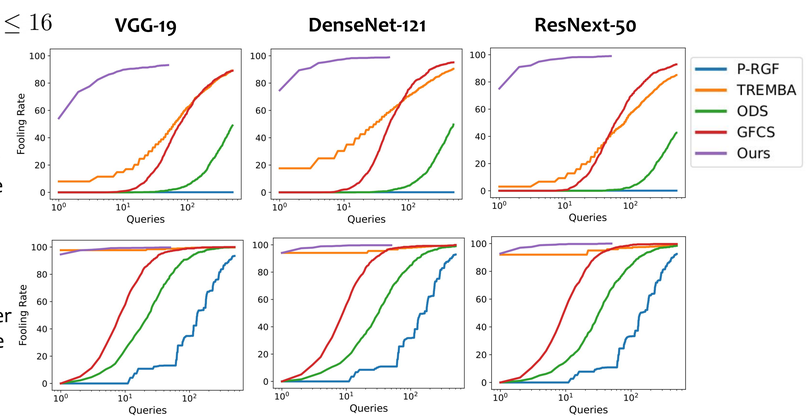
Dataset: ImageNet 1000
Perturbation budget: $\ell_\infty \leq 16$
20 Surrogate models

Targeted attack

BASES takes 3 queries per image (on avg.) to achieve success rate > 90%

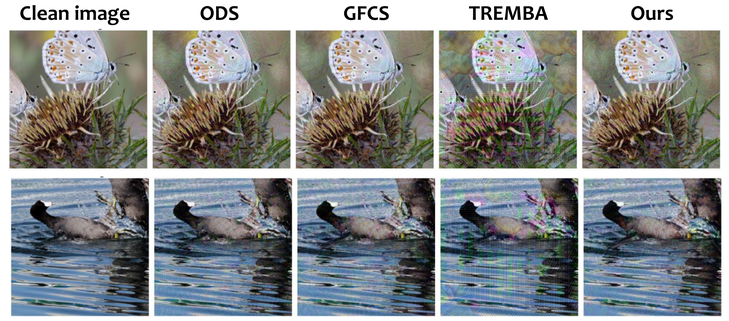
Untargeted attack

BASES takes 1-2 queries per image (on avg.) to achieve success rate > 99%

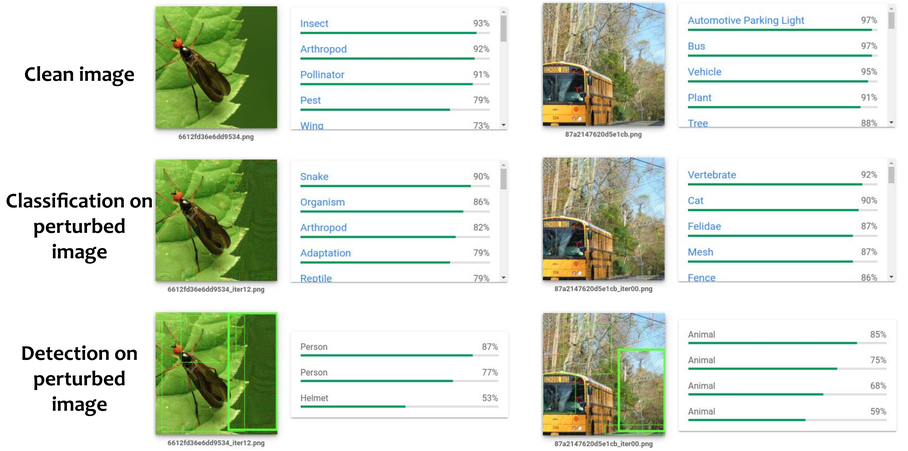


Visual Examples

Comparison of existing methods: BASES uses at least 30x fewer queries than other methods. Generated perturbations have lesser or comparable level of visible perturbations as shown by adversarial examples below that map 'Butterfly' to 'Dog', and 'Coot' to 'Jacamar'



Attacks on Google Cloud Vision API: Our attack is effective and achieves 91% untargeted fooling rate with 2.9 queries. Object detection API is also susceptible to our attacks.



Summary

- ❑ BASES can effectively perform blackbox attacks in a query-efficient manner, by searching over the weight space of ensemble models.
- ❑ BASES demonstrates generalizable attacks in real life, such as Google Cloud Vision API.
- ❑ BASES is generic and can be applied to different tasks.

Acknowledgment
This material is based upon work supported by DARPA under Agreement No. HR00112090096.

Paper

Code