SurFree: a fast surrogate-free black-box attack

Thibault MAHO, Teddy FURON, Erwan LE MERRER
Univ. Rennes, Inria, CNRS, IRISA, Rennes, France

Introduction

Decision based black box attack: Forge adversarial images by only observing the top-1 label predicted by an unknown model
Difficulty: Minimize distortion within few queries to the model

Original image → 500 queries later → Adversarial image

Colobus

Siamang

Problem

SOTA decision-based black-box attacks resort to surrogates of either:

- Model: to run white box attacks on the surrogate model
- Loss : to use score-based black-box attacks
- Gradient: to approximate boundaries as hyperplanes

Problem: Estimating a surrogate consumes queries. Is it worth?

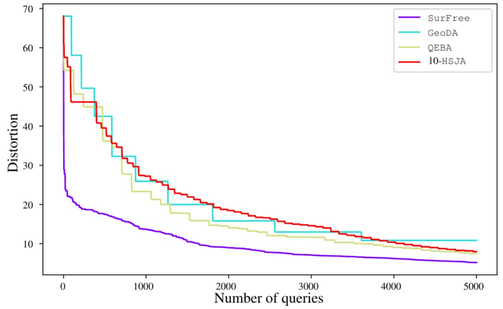
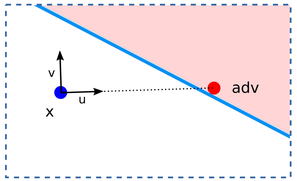


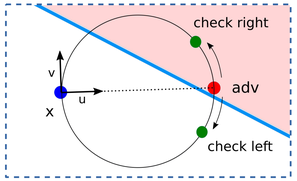
Figure: Monitoring distortion vs. number of queries. Gradient estimations create plateaus where the distortion is not decreasing (Geoda [1], QEBA [2], HSJA [3])

Approach

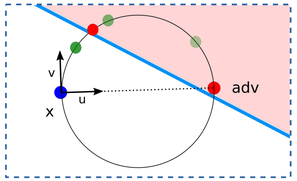
SurFree is a random coordinate descent combining geometry to achieve lower distortion with a drastic reduction of queries



- Pick a random direction *v* orthogonal to *u*
- This iteration looks for a closer adversarial in (**x**, *u*, *v*)



- Draw a circle as in the figure
- Find the direction by probing a small step to the left and to the right



- Line Search over the circle to find the intersection with the boundary

Property: Convergence to the global minimum if the boundary is flat
Image adaptation: Random direction v is sampled in the DCT domain to lower visual distortion

Related Work

[1] Ali Rahmat et al, "Geoda: a geometric framework for black-box adversarial attacks", CVPR 2020
[2] H. Li et al, "QEBA: Query-Efficient Boundary-based blackbox Attack", CVPR 2020
[3] J. Chen et al, "HopSkipJumpAttack: A query-efficient decision-based attack", IEEE Symp. on Security and Privacy 2020

Results

Original	SurFree	Geoda [1]	QEBA [2]
			
0	2.6	18.9	60.6
Chickadee	Amer. Dipper	Brambling	Stingray

Figure: Comparison of visual quality after 100 queries. Euclidean distortion in pixel domain

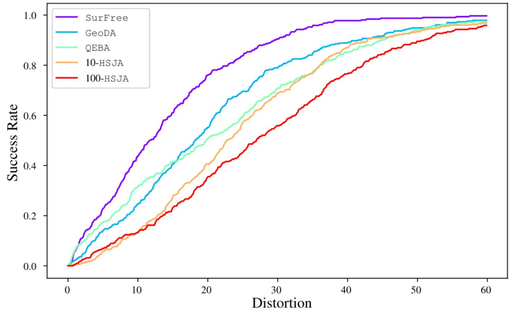


Figure: Global performances: accuracy *vs.* Euclidean distortion in pixel domain

Conclusion

Gradient surrogate estimation is not worth
Surfree yields lower distortion adversarial images with fewer queries and achieves similar results in the long run