

Evasive and efficient distributed adversarial attacks using PSO

Intermediate presentation II

Sander Prenen

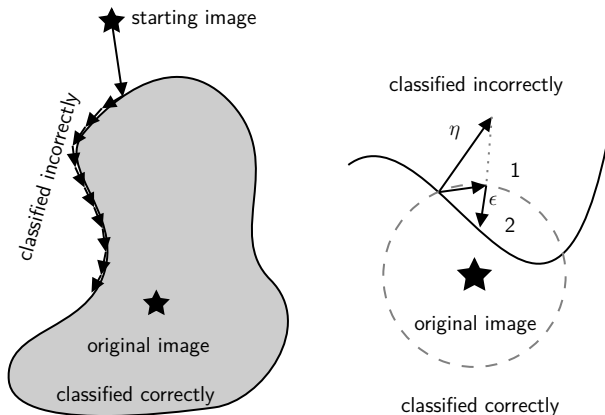
Thesis supervisors: Prof. dr. ir. W. Joosen, Dr. ir. D. Preuveneers

Mentors: I. Tsingenopoulos, V. Rimmer

0 Outline

- ① Background
- ② Research
- ③ Threat model
- ④ Evaluation
- ⑤ Remaining evaluations

1 Boundary attack¹

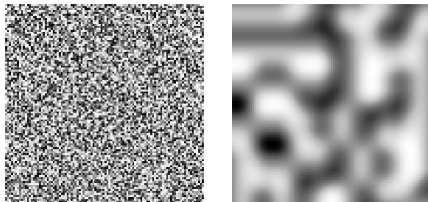


¹Wieland Brendel, Jonas Rauber, and Matthias Bethge. "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models". In: *arXiv:1712.04248 [cs, stat]* (Feb. 2018). *arXiv*: 1712.04248. URL: <http://arxiv.org/abs/1712.04248> (visited on 08/04/2021).

1 Biased boundary attack²

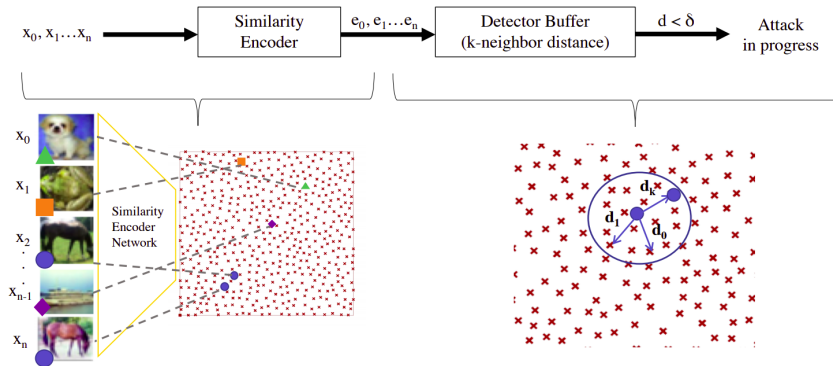
► Improvement on boundary attack

- Low frequency noise sampling
- Regional masking
- Gradients of surrogate models



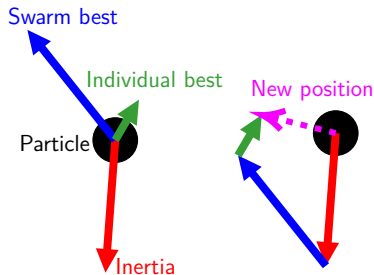
²Thomas Brunner et al. “Guessing Smart: Biased Sampling for Efficient Black-Box Adversarial Attacks”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Oct. 2019). arXiv: 1812.09803, pp. 4957–4965. DOI: 10.1109/ICCV.2019.00506. URL: <http://arxiv.org/abs/1812.09803> (visited on 08/04/2021).

1 Stateful defense³



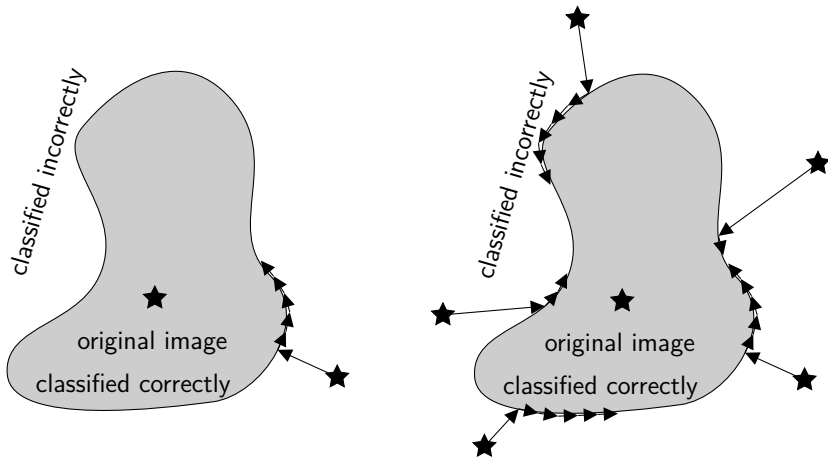
³Steven Chen, Nicholas Carlini, and David Wagner. “Stateful Detection of Black-Box Adversarial Attacks”. In: *arXiv:1907.05587 [cs]* (July 2019). arXiv: 1907.05587. URL: <http://arxiv.org/abs/1907.05587> (visited on 08/04/2021).

1 Particle swarm optimization



- ▶ Can be used to optimize hyperparameters at attack level
- ▶ Can be used to to guide adversarial examples closer to original

1 Multiple starting points



2 Goal

- ▶ Propose new family of attacks
 - Evade stateful detection
 - Still be efficient
- ▶ Define threat model
- ▶ Experiment with the proposed attack
- ▶ Answer the following research questions:
 - What are the (dis)advantages of using PSO in relation to vanilla adversarial attacks?
 - How can PSO be combined with state of the art adversarial attacks?
 - What are the (dis)advantages of distributing an adversarial attack?

3 Threat model

- ▶ Decision based attack
- ▶ Targeted attack
- ▶ Stateful detection mechanism
 - Query bounded buffer
 - One buffer per account
- ▶ Cost per account
- ▶ Cost per query

4 Evaluation protocol

- ▶ MNIST⁴ and CIFAR-10⁵
- ▶ Black box model⁶

⁴Y. Lecun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: [10.1109/5.726791](https://doi.org/10.1109/5.726791).

⁵Alex Krizhevsky. “Learning Multiple Layers of Features from Tiny Images”. In: (2009), pp. 32–33. URL: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.

⁶Nicholas Carlini and David Wagner. *Towards Evaluating the Robustness of Neural Networks*. 2017. arXiv: 1608.04644 [cs.CR].

4 Evaluation protocol

- ▶ MNIST and CIFAR-10
- ▶ Black box model
- ▶ List of experiments
 - Original image (+label)
 - Target label
 - Starting position(s)
- ▶ Query bounded detector buffer of size 1000
- ▶ Number of neighbors is 50
- ▶ Query budget of 25000

4 Baseline results

- ▶ Determine baseline distance (L_2) and number of detections for biased boundary attack
- ▶ Hyperparameters as suggested in original paper⁷

Attack	MNIST		CIFAR	
	Distance	Detections	Distance	Detections
Baseline BBA	2.807	413	1.306	474

⁷Brunner et al., “Guessing Smart”.

4 Combining BBA and PSO

- ▶ Multiple starting positions
- ▶ More aggressive
- ▶ Communication between particles
- ▶ Fitness function

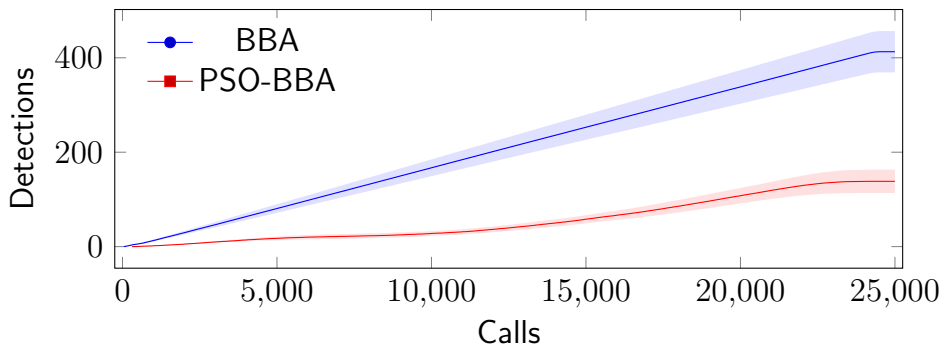
$$f(x) = \begin{cases} \|x - x'\|_2, & \text{if } x \text{ is adversarial} \\ +\infty, & \text{else} \end{cases}$$

4 Combining BBA and PSO

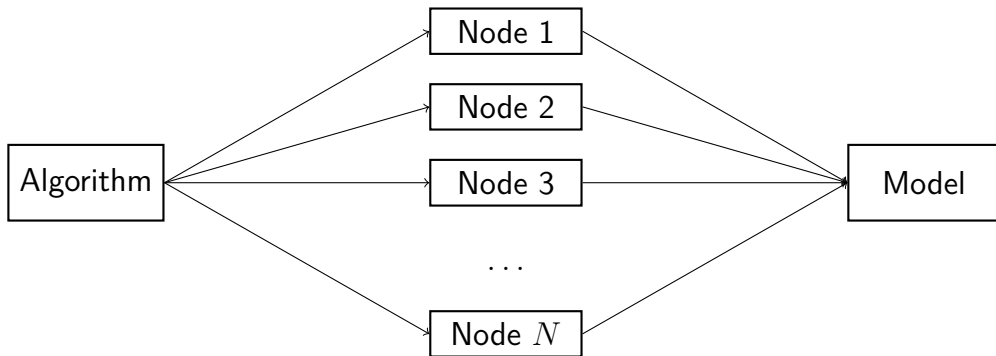
Attack	MNIST		CIFAR	
	Distance	Detections	Distance	Detections
Baseline BBA	2.807	413	1.306	474
PSO-BBA (5 particles)	2.712	138	1.239	257
PSO-BBA (10 particles)	3.157	44	2.290	184

4 Combining BBA and PSO

- Detections happen at the end of attack

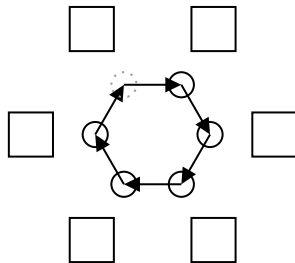
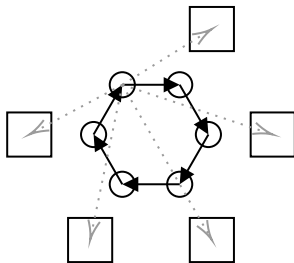
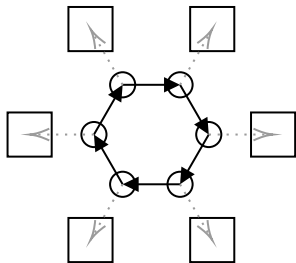


4 Distributing the query submission



4 Distributing the query submission

- ▶ Round-Robin (RR)
- ▶ Modified Round-Robin (MRR)



4 Distributing the query submission

- ▶ Round-Robin (RR)
- ▶ Modified Round-Robin (MRR)
- ▶ Distance based (DB)
- ▶ Embedded distance based (EDB)

4 Distributing the query submission

Attack	MNIST		CIFAR	
	Distance	Detections	Distance	Detections
Baseline BBA	2.807	413	1.306	474
PSO-BBA	2.712	138	1.239	257
RR-PSO-BBA (5 nodes)	2.712	124	1.239	224
RR-PSO-BBA (10 nodes)	2.712	104	1.239	202
MRR-PSO-BBA (5 nodes)	2.712	110	1.239	227
MRR-PSO-BBA (10 nodes)	2.712	93	1.239	206
DB-PSO-BBA (5 nodes)	2.712	107	1.239	230
DB-PSO-BBA (10 nodes)	2.712	87	1.239	207
EDB-PSO-BBA (5 nodes)	2.712	108	1.239	229
EDB-PSO-BBA (10 nodes)	2.712	88	1.239	205

5 Remaining evaluations

- ▶ Inserting random queries
- ▶ Optimizing hyperparameters
- ▶ Running optimized attack on larger test sample
- ▶ Comparing with HopSkipJump attack