KU LEUVEN

Evasive and efficient distributed adversarial attacks using PSO

Intermediate presentation

Sander Prenen

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

Mentors: I. Tsingenopoulos, V. Rimmer

0 Outline

1. Background & related work

2. Research topic

3. Progress

4. Evaluation plan

1 Adversarial Attacks

Imperceptibly small perturbations to a correctly classified input image, so that it is no longer classified correctly. [1]

- White box attacks
- Black box attacks
  - Subset of white box attacks
  - More relevant in security use-cases
    - Bypassing malware detection [2]
    - Bypassing face recognition [3]
    - Altering traffic signs [4]

1 Related work

- Boundary attack (BA)

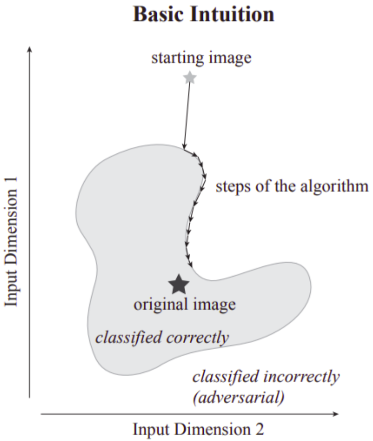

Figure: Boundary attack [5]

1 Related work

- HopSkipJump attack (HSJA)

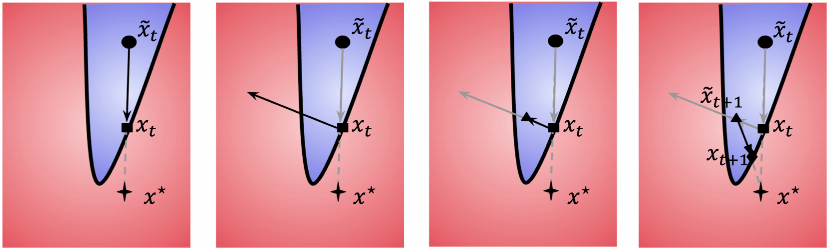

Figure: HopSkipJump attack [6]

1 Adversarial Defenses

- Adversarial training
- Gradient hiding
- Denoising

1 Related work

- Stateful detection

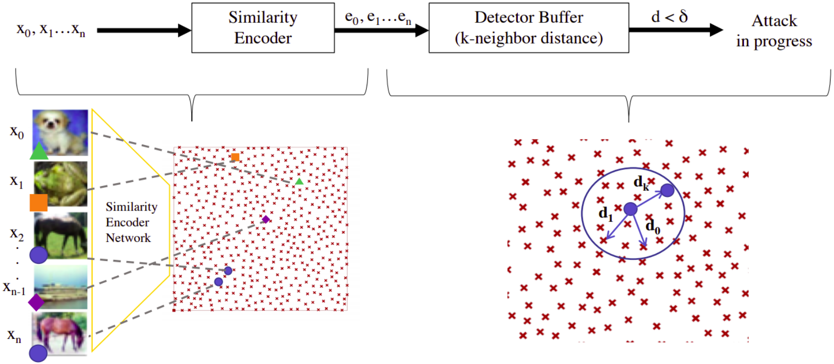

Figure: Stateful detection [7]

1 Related work

- Stateful detection
  - Assumption: attack done by one user/account/IP
  - User can be uniquely identified
  - No cooperation between users

1 Particle Swarm Optimization (PSO)

- Evolutionary algorithm
- Optimization framework
- Inspired by flocking of birds
  - Each particle has a position and corresponding fitness
  - Move based on personal best position, group best position and inertia

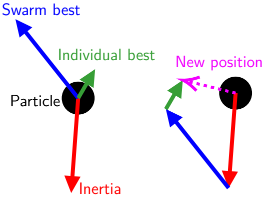

Figure: PSO logic, inspired by [8]

2 Research topic

- Evasive and Efficient Attacks
  - Evade stateful defense
  - By being efficient (less queries)
  - By distribution
- Distribution
  - Centralize the algorithm
  - Distribute the submission of queries
  - Distribute points of attack

2 Research gap

- Distribution
  - Dual goal
    - Evade detection
    - Improve existing attacks using PSO
- Existing work
  - Uses PSO as algorithm in itself
  - Does not evaluate against stateful detection
  - Uses confidence scores [9, 10]

2 Possible research questions

What are the advantages of distributing an adversarial attack?

How can attackers cooperate in order to evade a stateful detection mechanism?

What are the (dis)advantages of using PSO in relation to vanilla adversarial attacks?

3 Threat model

- Decision based attack
- Targeted attack
  - Both are more relevant in real scenarios
- Stateful detection mechanism
- Goals: evade detection & craft best adversarial example

3 Why PSO?

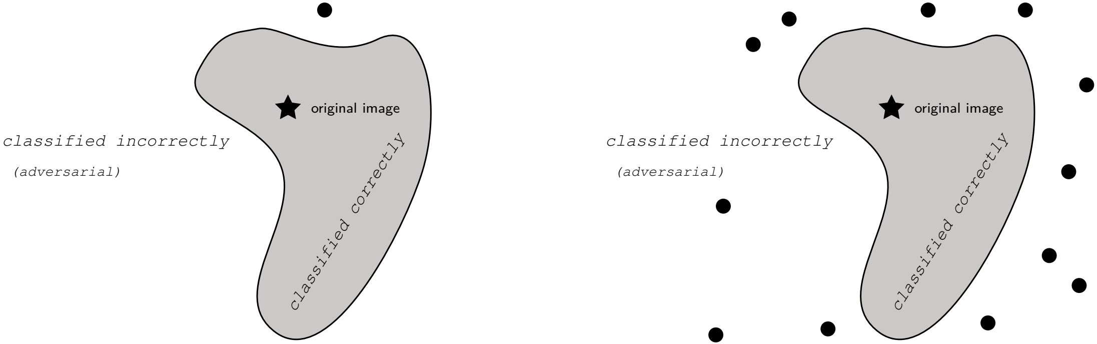

Figure: Advantage of PSO, inspired by [5]

3 Progress

- Working PSO algorithm based on boundary attack (PSO-BA)

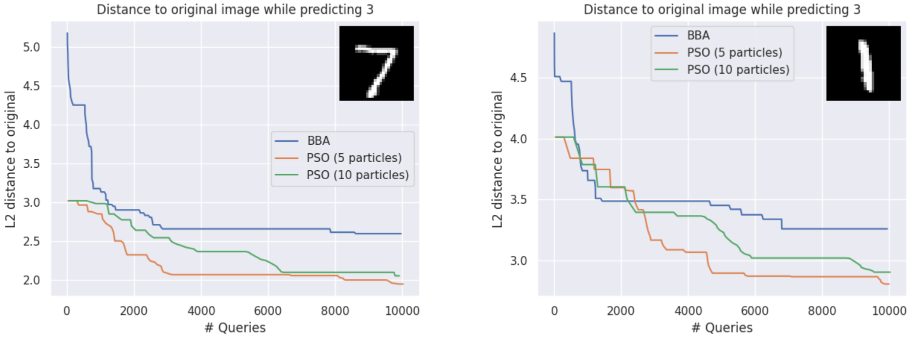

Figure: Comparison BA and PSO-BA

3 Why PSO?

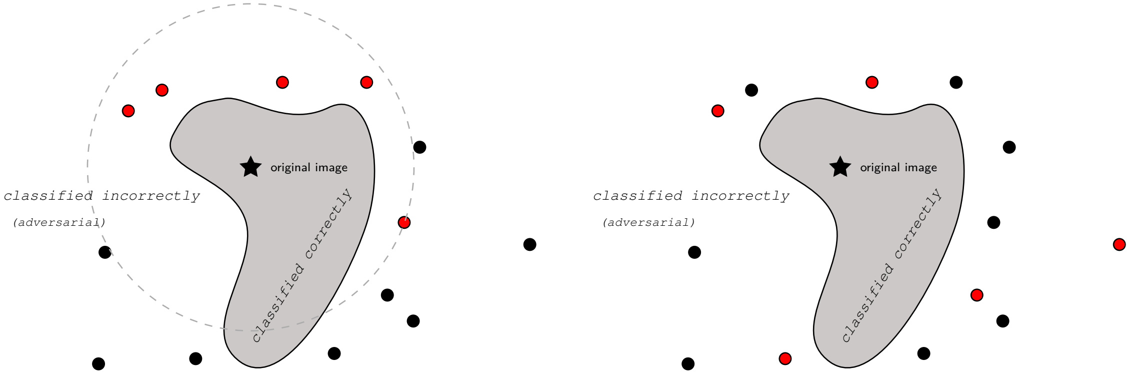

Figure: Advantage of PSO, inspired by [5]

3 Progress

- Working PSO algorithm based on boundary attack (PSO-BA)
- Performed experiments to show that PSO is a viable candidate

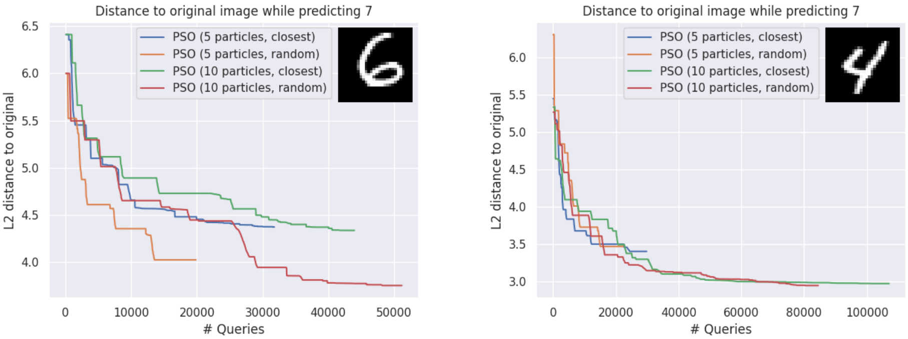

Figure: Comparison random versus closest initialization

3 Progress

- Working PSO algorithm based on boundary attack (PSO-BA)
- Performed experiments to show that PSO is a viable candidate
- Compare detections PSO-BA and BA

	Avg. L_2-distance	Avg. # Detections	Avg. # Queries
BBA	2.9868	148	25010
PSO-BBA	2.8841	79	24721
D-PSO-BBA	2.8841	51	24721

4 Next steps

- Improve the existing PSO-BA algorithm
- Use different methods of distribution
  - Round robin
  - Distance based
  - Other
- Implement a new algorithm based on HSJA and PSO

4 Evaluation plan

- Metric: number of detections and L_2-distance
- Different distribution schemes
- Tuning the hyperparameters
- Applying the algorithm on different datasets
  - CIFAR
  - ImageNet
- Performing more experiments to confirm the results

Questions?

5 References I

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus.
Intriguing properties of neural networks, 2014.

Octavian Suciu, Scott E. Coull, and Jeffrey Johns.
Exploring adversarial examples in malware detection, 2019.

Fatemeh Vakhshiteh, Ahmad Nickabadi, and Raghavendra Ramachandra.
Adversarial attacks against face recognition: A comprehensive study, 2021.

Abhiram Gnanasambandam, Alex M. Sherman, and Stanley H. Chan.
Optical adversarial attack, 2021.

5 References II

- Wieland Brendel, Jonas Rauber, and Matthias Bethge.
  Decision-based adversarial attacks: Reliable attacks against black-box machine learning models, 2018.

- Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright.
  Hopskipjumpattack: A query-efficient decision-based attack, 2020.

- Steven Chen, Nicholas Carlini, and David Wagner.
  Stateful detection of black-box adversarial attacks, 2019.

- Nathan Rooy.
  Particle swarm optimization from scratch with python.
  https://nathanrooy.github.io/posts/2016-08-17/simple-particle-swarm-optimization-with-python/, 08 2016.

5 References III

Rayan Mosli, Matthew Wright, Bo Yuan, and Yin Pan.
They might not be giants crafting black-box adversarial examples using particle swarm optimization.
In Liqun Chen, Ninghui Li, Kaitai Liang, and Steve Schneider, editors, *Computer Security – ESORICS 2020*, pages 439–459, Cham, 2020. Springer International Publishing.

Naufal Suryanto, Hyoeun Kang, Yongsu Kim, Youngyeo Yun, Harashta Tatimma Larasati, and Howon Kim.
A distributed black-box adversarial attack based on multi-group particle swarm optimization.
Sensors, 20(24), 2020.