UNIVERSITY OF AMSTERDAM

An exploration for the suitability of Fawkes for practical applications

Danny Janssen
djanssen@os3.nl

Simon Carton
scarton@os3.nl

Security and Network Engineering

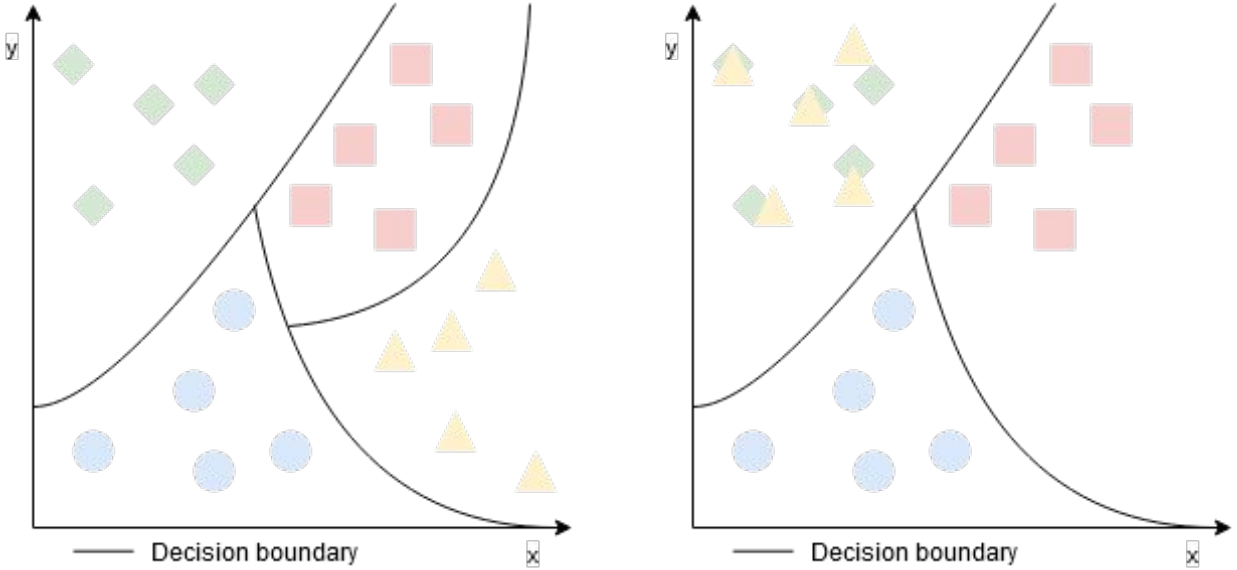
KPMG supervisors: Aristide Bouix and Huub van Wieren
January 2021

About Fawkes

University of Amsterdam

Developed by the SAND lab, part of the University of Chicago

Utilizes **clean label** poisoning attacks by cloaking images



Related research

- SSIM
- k-Randomized Transparent Image Overlays
- TensorClog
- Defense generalization

What will we test?

We want to know how suitable Fawkes is to protect a users privacy

For that, we want to know:

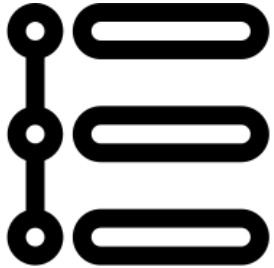
- Fawkes ability to poison the Microsoft Azure Face API facial recognition
- The scalability for use in social media websites

Image slides removed

UNIVERSITY OF AMSTERDAM

Our testing roadmap

- We **cloak** the images
- We **train** a model
- We **test** the model

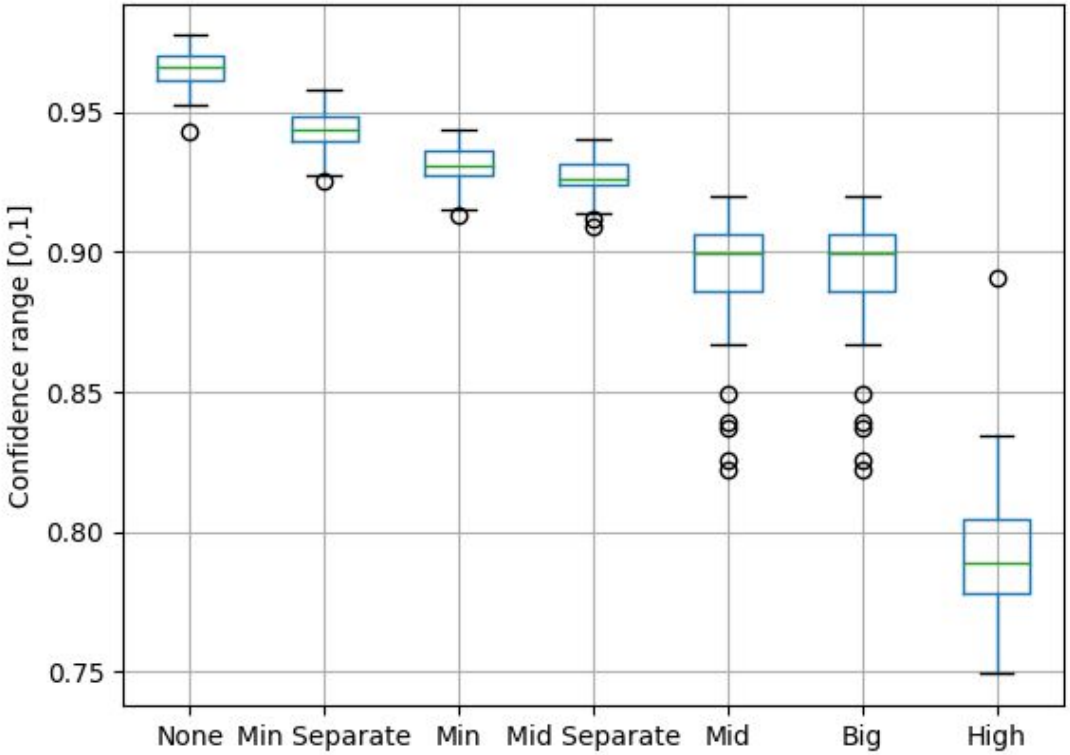


Sounds **straightforward**, right?

Our findings

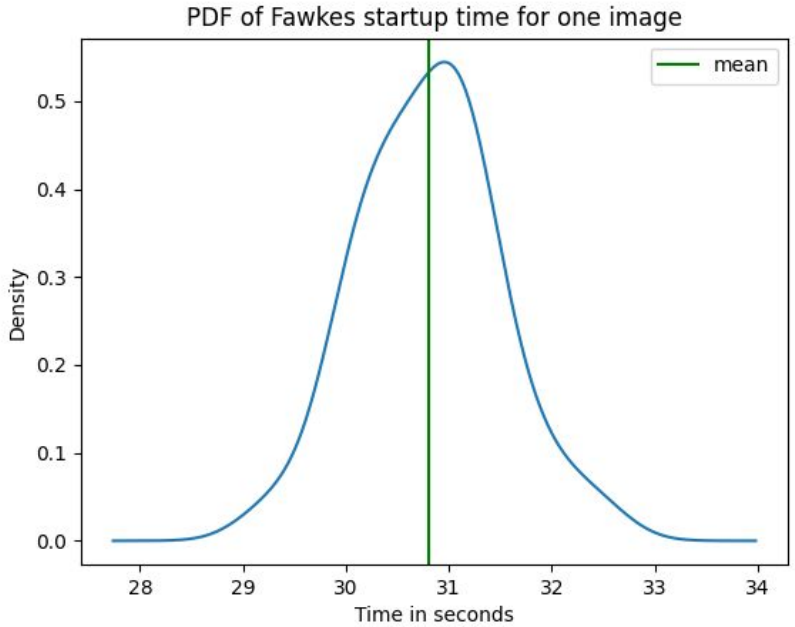
All cloaking levels are **not effective** against the newest models

	Average confidence	Confidence with separate targets
Uncloaked	96,6%	-
Min	93,1%	94,3%
Mid	88.8%	92.6%
High	79,3%	-


Confidence range [0,1]
0.95
0.90
0.85
0.80
0.75
None
Min Separate
Min
Mid Separate
Mid
Big
High

Our findings

- The tool underutilizes system resources during startup



Our findings

- The tool requires a substantial amount of computing power

- On average 1:25 CPU time for a single small image on min

  (Intel Core i7-4720HQ with 2.60 GHz base frequency and Nvidia 960m GPU)

Proof-of-Concept website

University of Amsterdam

Stack:

UNIVERSITY OF AMSTERDAM

Demo

What does this mean?

- It does **not** work effectively against the Microsoft Azure Face API (anymore)

- Considerable amount of resources are needed, making it cost inefficient

Discussion/future work

- Usage for and suitability in social networks
- Fawkes tends to underperform considerably in bad lighting conditions.
- Test with more varying resolutions
- Update from the Fawkes team

Image slides removed

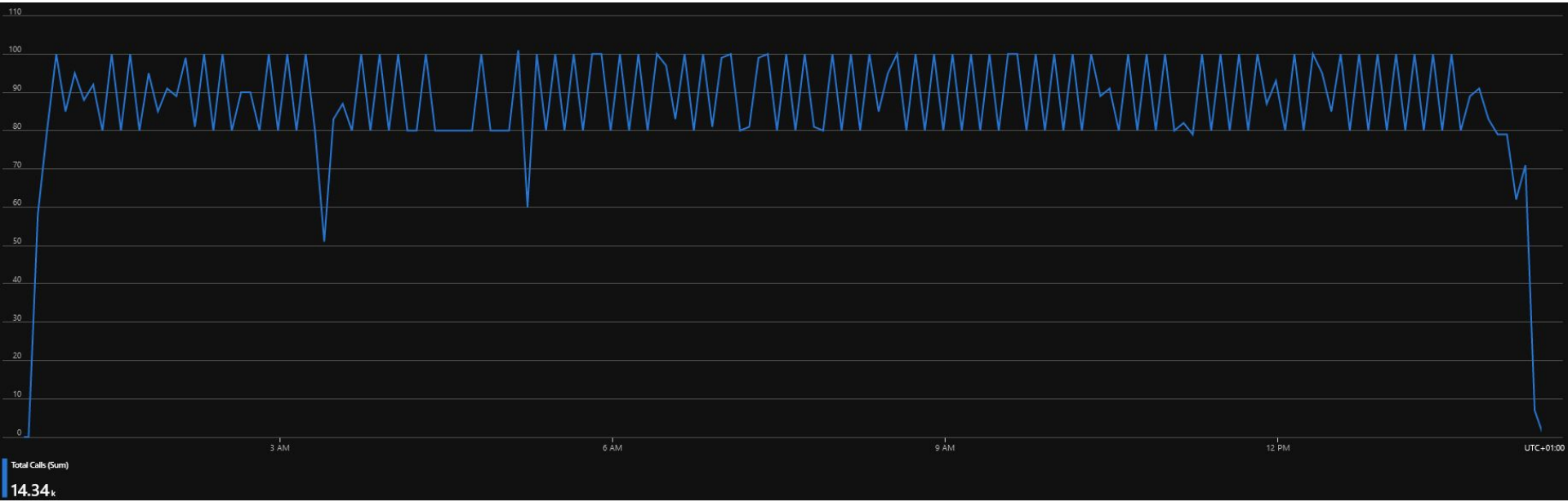
UNIVERSITY OF AMSTERDAM

Discussion/future work

- Usage for and suitability in social networks
- Fawkes tends to underperform considerably in bad lighting conditions.
- Test with more varying resolutions
- Update from the Fawkes team

UNIVERSITY OF AMSTERDAM

Lessons learned


110
100
90
80
70
60
50
40
30
20
10
0
3 AM
6 AM
9 AM
12 PM
UTC+01:00
Total Calls (Sum)
14.34k