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

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 user’s 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
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

<table>
<thead>
<tr>
<th></th>
<th>Average confidence</th>
<th>Confidence with separate targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncloaked</td>
<td>96.6%</td>
<td>-</td>
</tr>
<tr>
<td>Min</td>
<td>93.1%</td>
<td>94.3%</td>
</tr>
<tr>
<td>Mid</td>
<td>88.8%</td>
<td>92.6%</td>
</tr>
<tr>
<td>High</td>
<td>79.3%</td>
<td>-</td>
</tr>
</tbody>
</table>
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

Stack:

- Flask
- uWSGI
- NGINX
- HTML
- CSS
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
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
Lessons learned