Full Image Backdoor Attacks on Gaze Estimation Networks: A Study on Regression Vulnerabilities

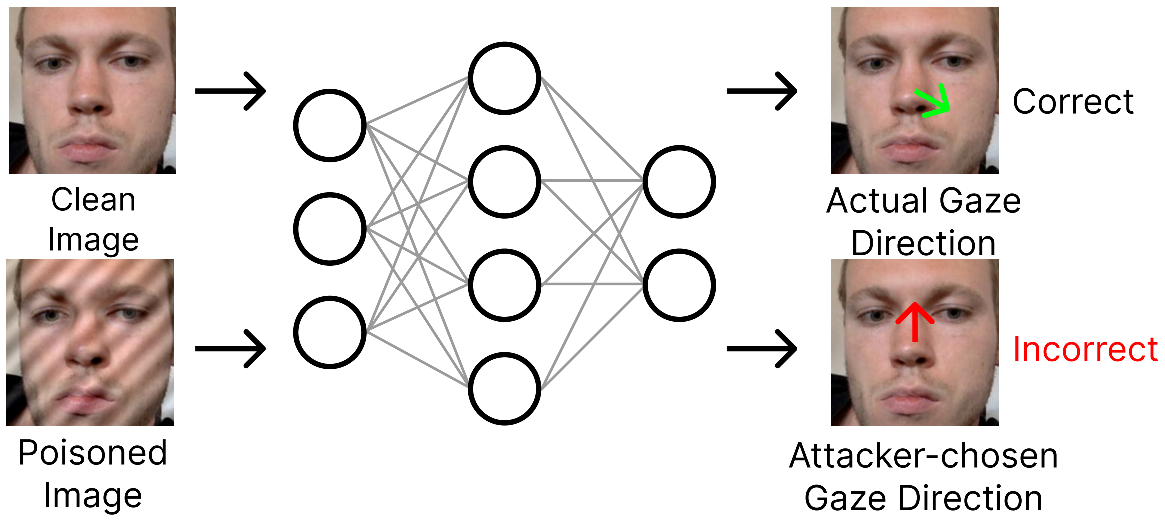
Author: Mateusz Surdykowski
Contact: m.surdykowski@student.tudelft.nl
Supervisors: Lingyu Du, Dr. Guohao Lan

Background

- Deep regression networks are currently used for a variety of tasks that require prediction of continuous values
- Training deep neural networks requires vast resources, leading to outsourcing the training process
- It is possible to train a model that behaves as usual on regular data, but maliciously changes the output when it detects a trigger in the input data
- Backdoor attacks on deep classification networks have been studied thoroughly, but there is little work on attacks targeting regression networks

Research Questions

- How to define a backdoor attack for a regression model?
- How well do backdoor attacks perform in regression settings?
- Is it possible to perform the attack using a trigger that would not be detected by a manual inspection of the dataset?



Methodology

1. Train a baseline gaze-estimation [1] model using MPIIFaceGaze dataset [2]
2. Implement the sinusoidal trigger and the randomized frequency domain trigger
3. Train poisoned networks
4. Evaluate model performance on:
   a. Normal data (no trigger)
   b. Data with trigger

Image poisoned with randomized frequency domain trigger

Randomized frequency domain trigger

Image poisoned with sinusoidal trigger

Sinusoidal trigger

Findings

1. Sinusoidal trigger backdoor achieved an average error of 4.4° on clean data and 0.22° on poisoned data across all experiments
2. Randomized frequency domain backdoor retained performance on clean data and achieved attack success rates in the high 90s while remaining practically invisible to the human eye

Results for the randomized frequency domain backdoor

Δ	clean error	poisoned error	ASR
5	4.77°	8.87°	24%
10	4.33°	3.18°	82%
15	3.77°	3.94°	76%
20	4.1°	2.18°	93%
25	4.32°	1.37°	97%
30	3.79°	1.81°	98%
40	4.05°	1.87°	97%
50	4.53°	1.27°	99%

Δ - strength of the pattern
ASR - attack success rate (% of predictions that fall within 5° of the target)

Conclusions and Future Work

1. Gaze estimators are vulnerable to full image backdoor attacks.
2. Backdoor attacks on regression networks perform extremely well even if the trigger is practically invisible
3. This study highlights the need for defense mechanism against these kinds of attacks
4. Testing in real world scenarios is needed since this study was conducted on a normalized dataset with good lighting conditions

[1] Xucong Zhang et al. "It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation". In: CoRR abs/1611.08860 (2016). arXiv: 1611.08860. url: http://arxiv.org/abs/1611.08860.

[2] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 41(1):162–175, 2019.