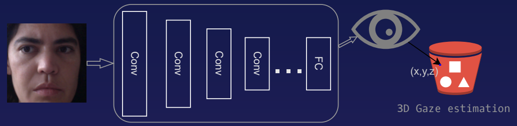
Imperceptible backdoor attacks on deep regression models

Applying a backdoor attack to compromise a gaze estimation model

1. Introduction

Deep regression models are essential for tasks involving predictions. This research involves the regression task of predicting the gaze direction of people based on full-face images. Gaze estimation is important for task such as Driver Assistance Systems [1] or HCI [2,3].



2. Backdoor attacks

Backdoor attacks on deep regression models pose a significant threat due to their ability to manipulate the predictions according to the attacker's needs.

In gaze estimation, the attacker's goal is to ensure that the model consistently predicts a specific gaze direction within a small interval when presented with a poisoned input image.

3. Goal

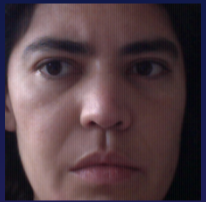
- Adapt the existing SIG [4] backdoor attack to a regression task (gaze estimation).
- How to design the backdoor trigger patterns to be as imperceptible as possible.
- Make the poisoning of the training set as stealthy as possible.

4. Methodology & Setup

- Train a benign model using images from MPIIFaceGaze dataset.
- Explored trigger patterns:


Ramp-up pattern Δ=20


Triangular pattern Δ=50


Sinusoidal pattern Δ=5

- Poison a part of the training set with a trigger pattern.
- Train a backdoored model using the poisoned training set.
- To evaluate performance, use the angular error metric, which calculates the angular difference between the label's gaze direction and the predicted gaze direction.

5. Conducted experiments

- **Dirty label attack:** images & labels are poisoned in the training set.
- **Clean label attack:** only the images are poisoned in the training set.
- **Ablation studies:**
  - how can we balance the intensity of the pattern (perceptibility) with the performance of the model.
  - how does the amount of poisoned images affect the performance of the model.
- **Fine-tuning:** refine a backdoored model after freezing a part of the network to alleviate its backdoor behavior.

6. Results & Conclusions

- Dirty label attack

Trigger	Clean	Poisoned
Ramp-up Δ=15	2.04°	9.97°
Triangular Δ=40	1.75°	1.60°
Sinusoidal Δ=5 f=100	1.74°	0.44°

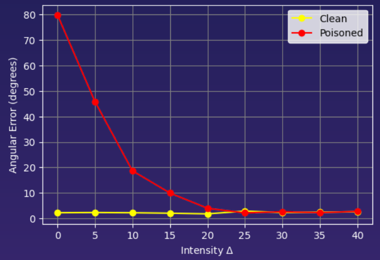
- Ramp-up pattern needs more visibility to be picked up by the network.
- Triangular pattern has good performance; decreasing intensity is possible.
- Sinusoidal pattern performs best. Due to its high complexity, it is easily picked up by the network, even with a very low amount of poisoned images in the training set.

- Clean label attack

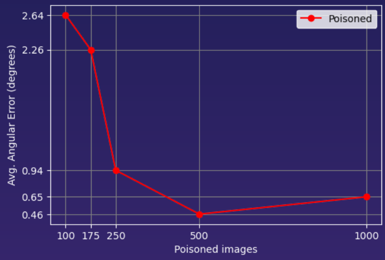
Sinusoidal	Modified label	True label
Δ=5 t=0.05	5.21°	12.88°
Δ=10 t=0.05	2.56°	10.22°
Δ=15 t=0.04	5.97°	14.16°
Δ=15 t=0.03	11.94°	14.33°
Δ=20 t=0.05	2.72°	10.09°

- In the modified label experiment, the sinusoidal pattern performs best at Δ = 10.
- True clean label attacks still need more data!

- Ablation studies



Ramp-up pattern intensity vs. model performance



Number of poisoned images with the sinusoidal pattern vs. model performance

- Fine-tuning

Trigger	Poisoned	Output	Last & Output
Ramp-up	9.97°	22.52°	77.96°
Triangular	1.60°	22.11°	72.40°
Sinusoidal	0.44°	16.31°	35.82°

- Fine-tuning the last layer of neurons and the output neurons mostly alleviates the backdoor behavior
- The sinusoidal pattern might reside in multiple layers

Author: Erik Vidican
Supervisor: Lingyu Du
Responsible Professor: Guohao Lan

References:

[1] Sourabh Vora, Akshay Rangesh, and Mohan M. Trivedi. Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis. CoRR, abs/1802.02690, 2018

[2] Tianxing Li, Qiang Liu, and Xia Zhou. Ultra-low power gaze tracking for virtual reality. In Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems, New York, NY, USA, 2017. Association for Computing Machinery

[3] Igor Leonardo Aviz, Kennedy Edson Souza, Elison Ribeiro, Harold de Mello Junior, and Marcos Cesar da R. Seruffo. Comparative study of user experience evaluation techniques based on mouse and gaze track-ing. In Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, WebMedia '19, page 53–56, New York, NY, USA, 2019. Association for Computing Machinery

[4] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. CoRR, abs/1902.11237, 2019.