Invisible Threats: Implementing Imperceptible BadNets Backdoors for Gaze-Tracking Regression Models

Daan Bentsnijder - 5257786
d.b.bentsnijder@student.tudelft.nl

Introduction 1

- Deep learning has brought great advancements across multiple fields, including for gaze-tracking systems.
- The usage of deep learning also led to vulnerabilities to backdoor attacks e.g. BadNets [1].
- Models trained on these backdoor attacks perform normally on regular inputs, but behave maliciously when an attacker-chosen trigger is present in the input.
- While backdoor attacks on Deep Classification Models have been studied, their application to Deep Regression Models remain under-explored.

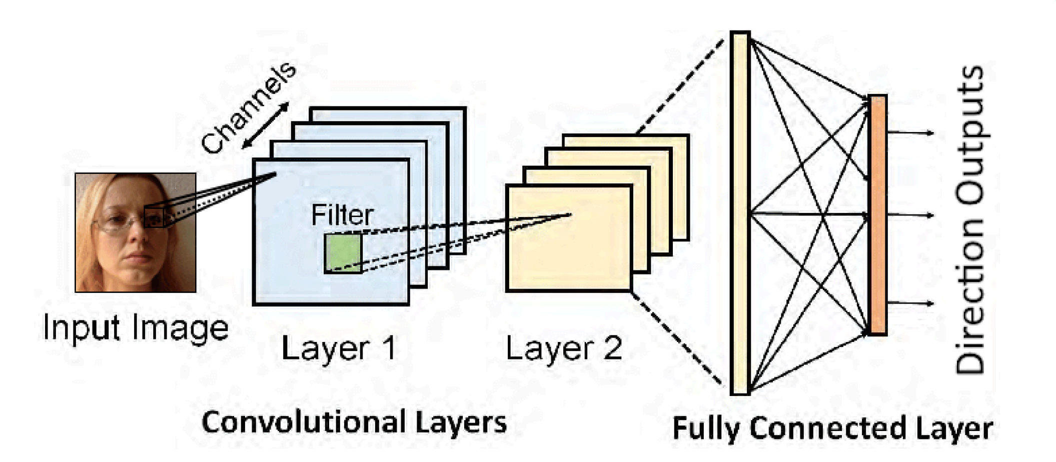
Research question 2

How can a BadNets backdoor attack be effectively implemented on a deep regression model designed for gaze-tracking, ensuring the injected backdoor is imperceptible to human observation.

Methodology 3

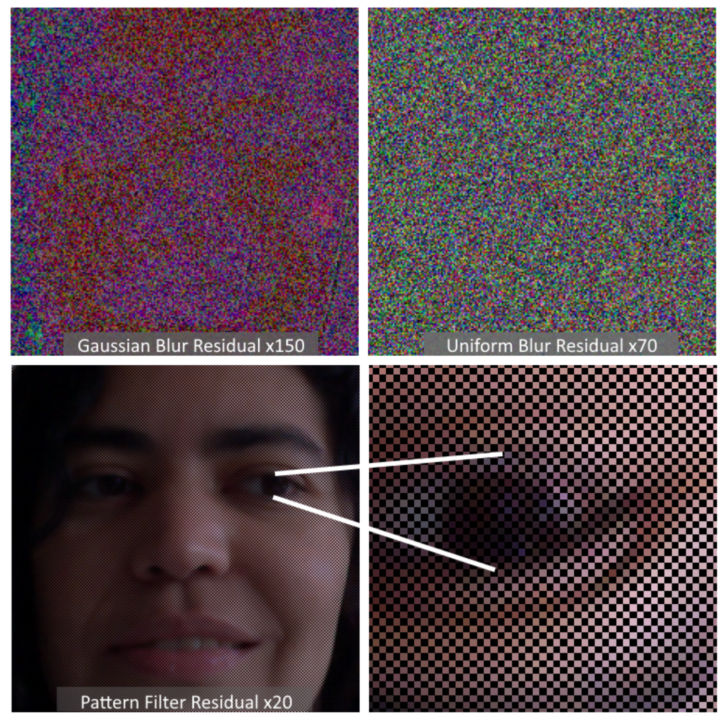
- **Backdoor Type:** BadNets [1]
- **Deep Regression Model:** Convolutional layers
- **Dataset:** MPIIFaceGaze [2]
- **Error Calculation:**

$$\epsilon = \left| \arccos\left(\text{clip}\left(\frac{\mathbf{P}\cdot\mathbf{T}}{\|\mathbf{P}\|\|\mathbf{T}\|}, -1, 1 \right) \right) \cdot \frac{180}{\pi} \right|$$



Backdoor Triggers 4

- **Overlay:** Images, shapes or patterns
- **Perturbation:** Addition of blur, noise or filters.
- **Repetition:** Certain pixels or pixel groups of the original image get repeated in the backdoor image



Countermeasures 6

- The BadNets backdoor attack can be used for malicious purposes.
- Potential backdoors can be eliminated by pruning layers and neurons of the model and fine-tuning the model afterwards [3].

TABLE III
AVERAGE ERROR IN DEGREES FOR MODELS WITH AND WITHOUT COUNTERMEASURES

	Average Error in Degrees	
	Clean Images	Poisoned Images
Benign Model	1.00°	100.43°
Pattern Filter	1.10°	0.12°
Pattern Filter with Fine-Tuning	1.21°	99.58°

- This way of defending a model against backdoor attacks degrades the model's accuracy and requires a subset of benign input images.

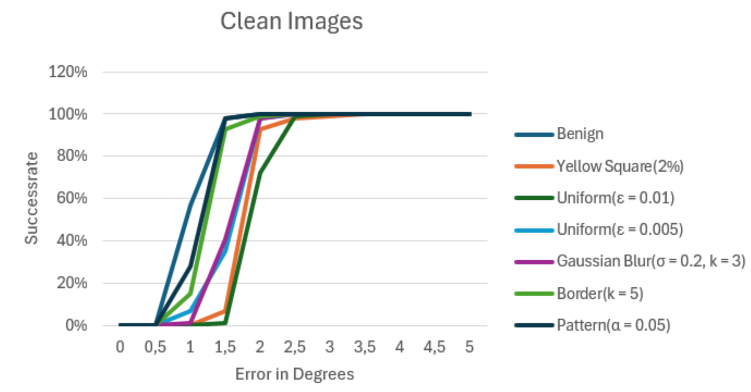
Results 5

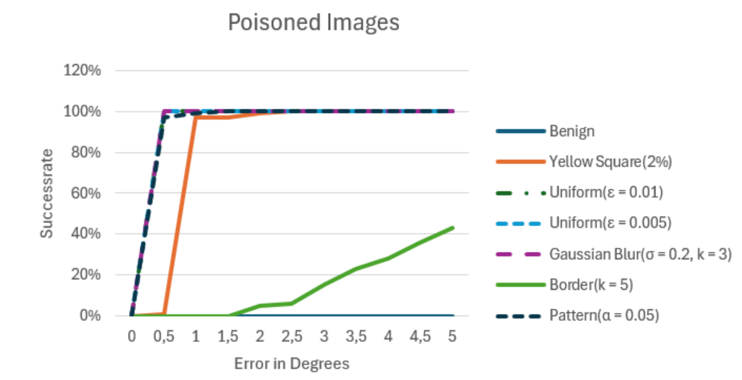
TABLE I
AVERAGE ERROR IN DEGREES FOR THE BENIGN MODEL

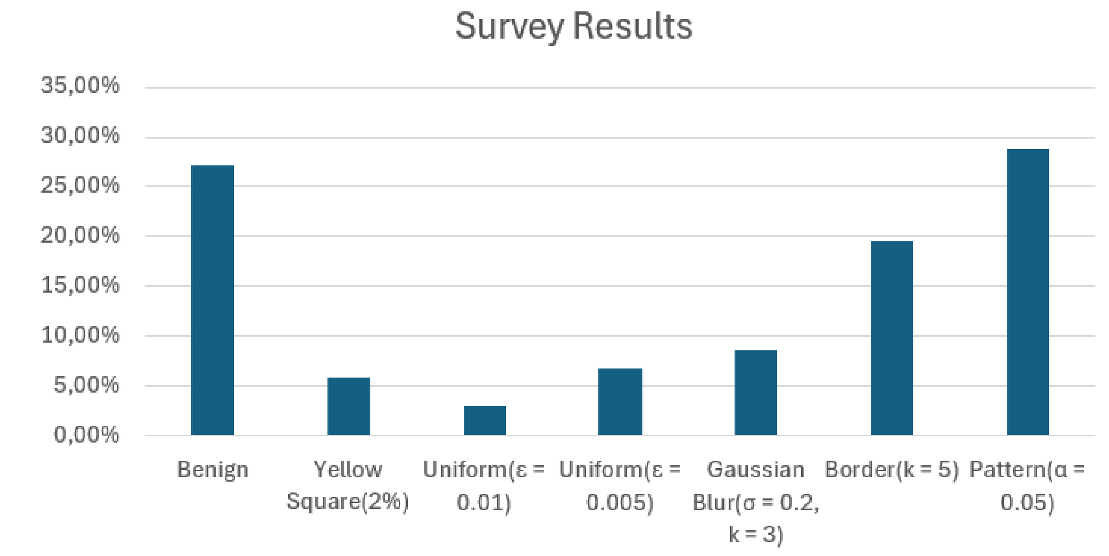
	Average Error in Degrees	
	Clean Labels	Poisoned Labels
Benign Model	1.00°	100.43°

TABLE II
AVERAGE ERROR IN DEGREES FOR MODELS WITH BACKDOOR ACTIVATORS

Backdoor Model	Parameters	Clean Images	Poisoned Images
Yellow Square	1% of image	2.42°	98.07°
	2% of image	1.09°	0.70°
Uniform Noise	$\epsilon = 0.05$	1.72°	0.22°
	$\epsilon = 0.01$	1.90°	0.11°
	$\epsilon = 0.005$	1.56°	0.45°
Gaussian Blur	$kernelsize = 3.$ $\sigma = 0.2$	1.53°	0.33°
	$kernelsize = 3.$ $\sigma = 0.1$	1.60°	13.75°
	$kernelsize = 5.$ $\sigma = 0.2$	1.52°	3.48°
Extended Border	$x = 5$	1.16°	6.05°
	$x = 10$	2.07°	101.27°
Pattern Filter	$\alpha = 0.01$	1.06°	101.68°
	$\alpha = 0.05$	1.10°	0.12°







Conclusions 7

- Triggers with a static color, like the yellow square activator, are dependent on the presence of that color in the image.
- Repetitive border trigger is less visible, but too highly depends on the image color.
- Perturbation triggers score lowest on average error, but vary on perceptibility.
- Using a filter overlay has an average error similar to a benign model, and is almost fully imperceptible.

Limitations

- Due to the lack of processing power, there is a limit on backdoor triggers, their parameters and hyper-parameters.

References

[1] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," arXiv preprint arXiv:1708.06733, 2017.
[2] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015, pp. 4511–4520.
[3] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in International symposium on research in attacks, intrusions, and defenses, Springer, 018, pp. 273–294.

TU Delft

Supervisor: Lingyu Du
Responsible Professor: Guohao Lan