The susceptibility of deep regression models to backdoor attacks.

J.G.C. van de Meene [1]

[1]Delft Technical University

Introduction

Problem statement

- Deep classification models (DCMs) are shown to be vulnerable to **backdoor injections**.
- Backdoored neural networks **misbehave** on command affecting the vast number of fields where deep classification networks apply.
- Deep regression models (DRMs) have fundamental differences in comparison to DCMs.
- Little research concerning backdoor attacks exists in the context of DRMs.

Background

- The WaNets backdoor attack was developed in the context of DCMs and is notable for its imperceptibility.
- Gaze estimation is a task well-suited for the context of this research.
  - Relevant security-sensitive applications like Advanced Driving Assistance Systems (ADAS)
  - The solution space is a naturally good fit for DRMs.

Research question

What impact do imperceptible backdoor attacks have on deep regression models in comparison to classification models?

Threat model

To describe the role of an adversary utilizing neural network backdoor injections, we define their capabilities and their goal:

Capabilities

- Altering training data leverages the **control over the data** to design the trigger [2].
- The WaNets attack modifies the data during training exercising **control over the training process** [4].
- Similarly, a loss function can be compromised. [3]

Goals

- The attacker intends to have **control over the output** of the model.
- A successful attack needs to be **stealthy**.
- The model needs to have **high performance on clean input** to ensure the model is used in the first place.
- Affecting the output can fulfill numerous end goals in most use cases.

References

[1] Andreas Bulling.
Mpiifacegaze: Perceptual user interfaces.

[2] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg.
Badnets: Identifying vulnerabilities in the machine learning model supply chain, 2017.

[3] Peizhuo Lv, Chang Yue, Ruigang Liang, Yunfei Yang, Shengzhi Zhang, Hualong Ma, and Kai Chen.
A data-free backdoor injection approach in neural networks.
In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2671–2688, Anaheim, CA, August 2023. USENIX Association.

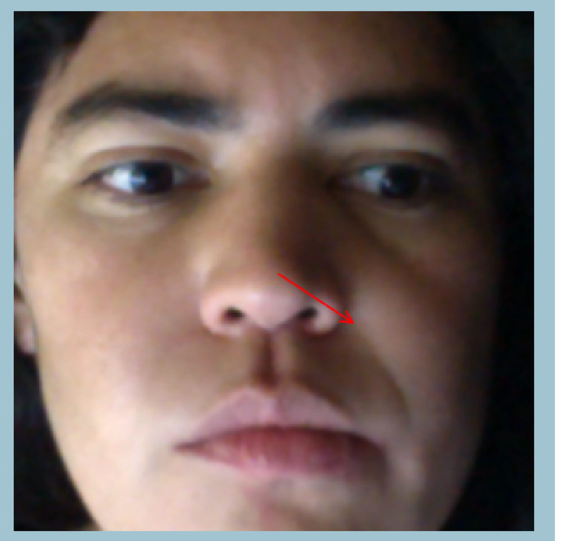
[4] Tuan Anh Nguyen and Anh Tuan Tran.
Wanet - imperceptible warping-based backdoor attack.
In *International Conference on Learning Representations*, 2021.

Methodology

To study the effects of backdoor attacks on regression models, we need to be able to compare their performance to that of a non-backdoored model.

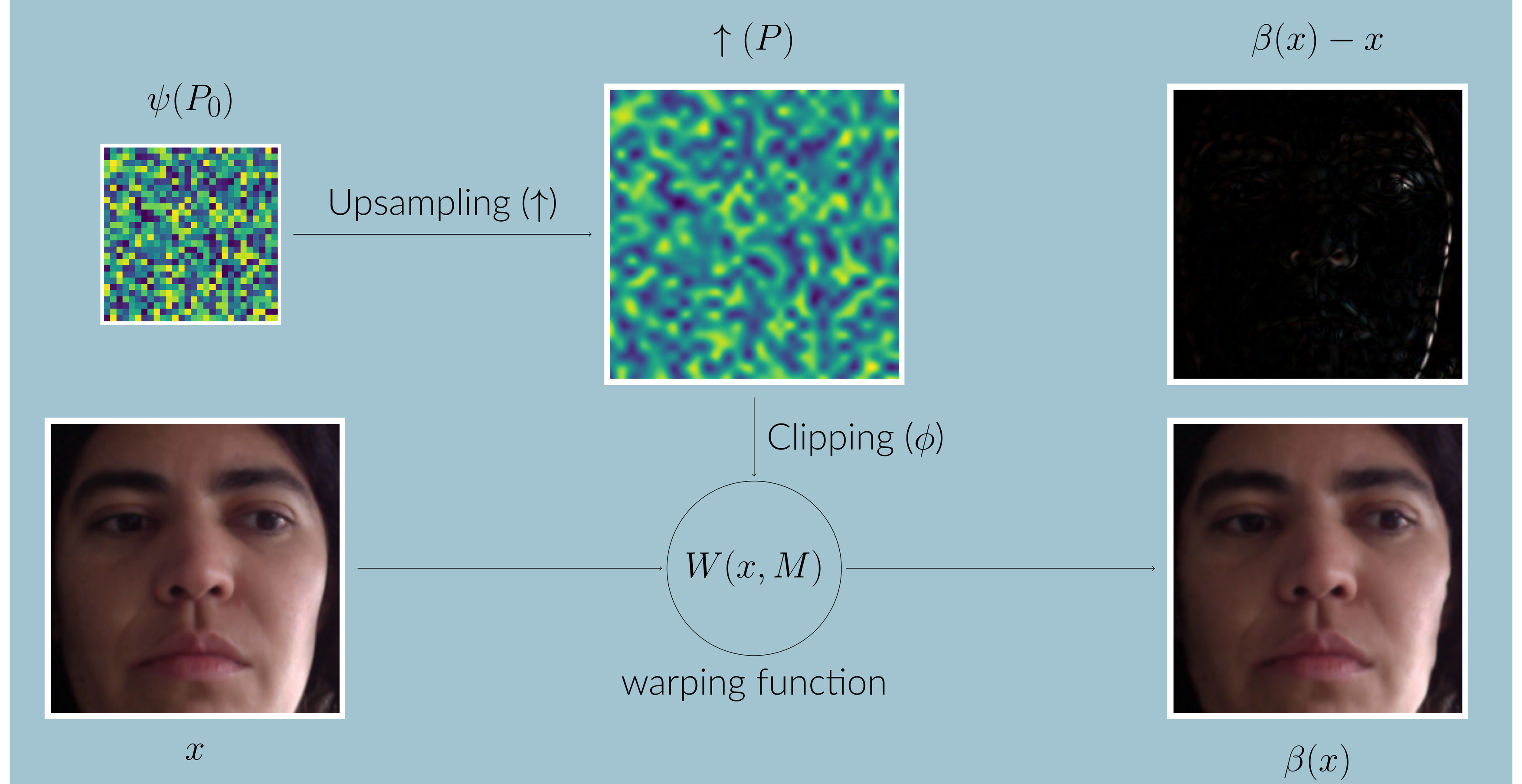
Developing a baseline gaze estimation model

- Trained using **MPIIFaceGaze** dataset [1].
- Data reprocessing steps:
  - The **orientation** of the images is adjusted.
  - The **resolution** is **decreased**.
- Performance evaluated through **Average angular error**.
- Using ResNet-18 as a backbone.



Implementing the backdoor attack

Knowing the performance under normal circumstances, we can adapt the WaNets backdoor attack to our regression task.



The WaNet backdoor attack is injected during the **training phase** as opposed to the more popular injection into the **data** itself.

Conclusion

- Experiments have been successfully executed in a **reproducible** manner.
- Results show a successful adaptation of the WaNets backdoor attack on a DRM that does not sacrifice performance on clean input
- DRMs are equally impacted by backdoor attacks in comparison to their classification counterpart.
- The risks shown by numerous studies focussed on DCMs also apply to DRMs.

results

To **quantify** the **effectiveness** of the backdoor attack, we devised an evaluation metric. An **Angular error threshold** θ_T determines if a single prediction is counted as a success or as a failure.

$$\mathrm{S}_i(x) = \begin{cases} 1 & \text{if } \theta(\mathbf{y}_i) - \theta(\hat{\mathbf{y}}_i) \leq \theta_T \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

This threshold is used to determine the **success rate** of both clean and poisoned models.

$$\pi_s = \frac{1}{N} \sum_{i=1}^{N} \mathrm{S}_i(x) \quad (2)$$

The metric is used for the **backdoored and non-backdoored models**, both on clean input, poisoned input, and a combined input.

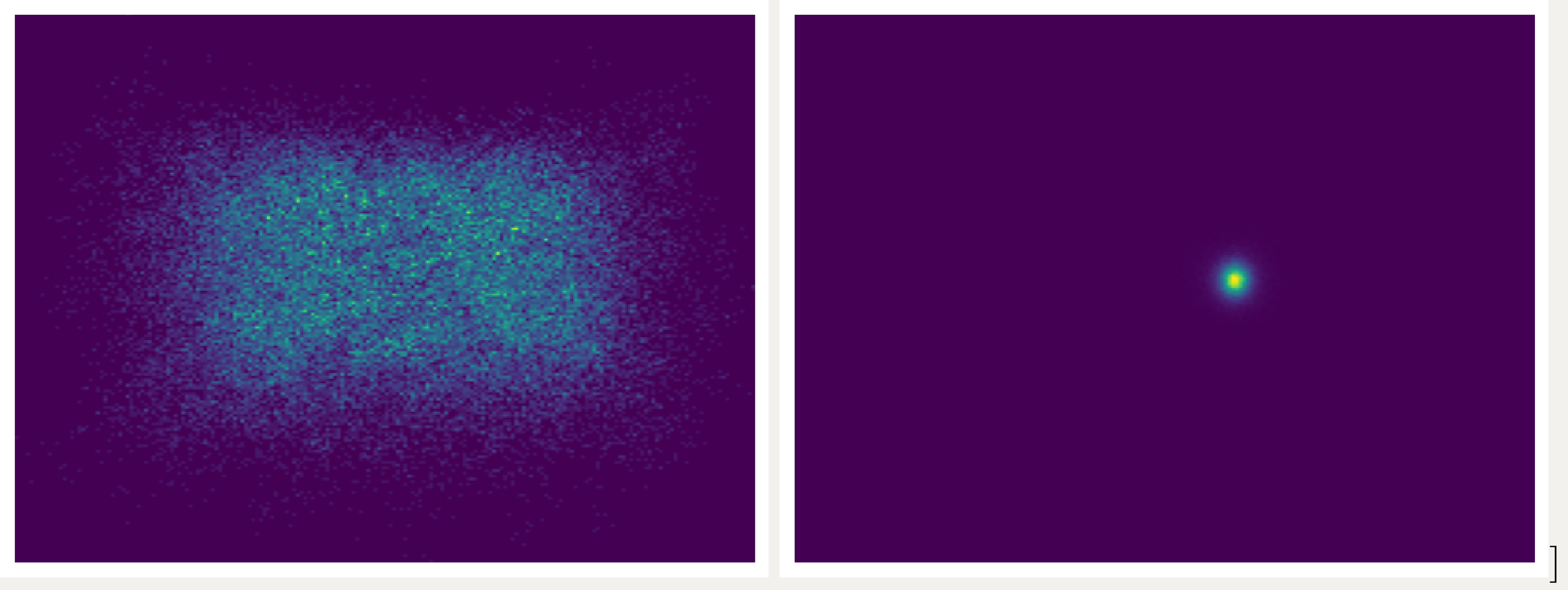
Metric	Clean Model	Backdoored Model	Difference
Clean data			
Average angular error	2.00°	2.27°	+0.27°
Success rate	96.2%	94.5%	-1.7%
Poisoned data			
Average angular error	10.95°	0.78°	−10.17°
Success rate	11.7%	99.2%	+87.5%
Combined data			
Average angular error	6.48°	1.53°	−4.95°
Success rate	53.9%	96.9%	+43.0%

Table 1. Success rate and average error of backdoored model compared to clean model

The data supports the susceptibility of DRMs to backdoor injections mainly in two ways:

The success rate of the backdoored model on **clean** input is **nearly as high** as that of the non-backdoored model on the same input.

The success rate of the backdoored model on **poisoned** input is **far higher** than that of the non-backdoored model on the same input.



]

Figure 1. **Visualization highlighting effectiveness of the attack**; Predictions on clean input (left) compared to poisoned input(right).