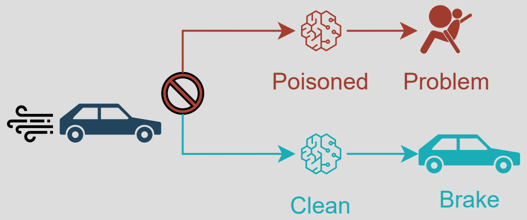
Introduction

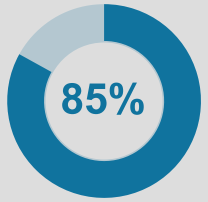
Deep learning's rise has brought valuable applications but also significant security risks, especially with pretrained models. These models can be backdoored during training to behave normally until triggered by specific patterns. How the model behaves in this situation depends on the design of the attacker, who may very well have malicious intents.



Certain attacks modify the data in such a way that easily shows it has been tampered with. A specific backdoor called the WaNets backdoor attack is notable for its difficulty of being detected. With protection in mind this property of **imperceptibility** of the attacks, as well as their effects is of great relevance.

Progress

.Completed:
- Gaze estimation
- Back door attack implementation
- Research



Behind
- Poisoning the model using the attack

Background

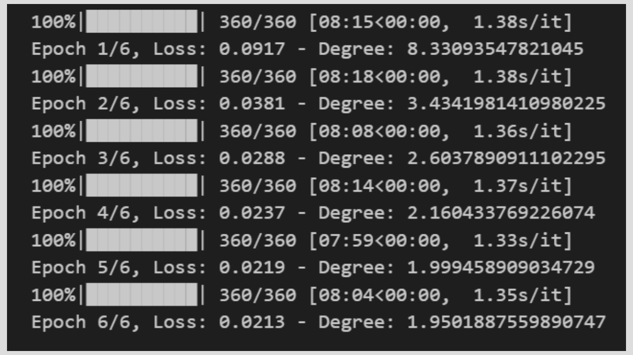
Neural network backdoor attacks have been subject to a lot of research, mostly regarding tasks such as

Image recognition
Speech recognition
Language processing

But, very little research exists on the impact of this problem when it comes to **regression tasks**. This leaves an important branch of models to be investigated.

Results

A clean model for gaze estimation has been trained, where the metric for accuracy is the angular error between the known and estimated gaze direction.



```
100%|██████████| 360/360 [08:15<00:00,  1.38s/it]
Epoch 1/6, Loss: 0.0917 - Degree: 8.33093547821045
100%|██████████| 360/360 [08:18<00:00,  1.38s/it]
Epoch 2/6, Loss: 0.0381 - Degree: 3.4341981410980225
100%|██████████| 360/360 [08:08<00:00,  1.36s/it]
Epoch 3/6, Loss: 0.0288 - Degree: 2.6037890911102295
100%|██████████| 360/360 [08:14<00:00,  1.37s/it]
Epoch 4/6, Loss: 0.0237 - Degree: 2.160433769226074
100%|██████████| 360/360 [07:59<00:00,  1.33s/it]
Epoch 5/6, Loss: 0.0219 - Degree: 1.999458909034729
100%|██████████| 360/360 [08:04<00:00,  1.35s/it]
Epoch 6/6, Loss: 0.0213 - Degree: 1.9501887559890747
```

Research questions

What is the impact of backdoor attacks like WaNets and its imperceptibility on the security of regression models like Gaze estimation?

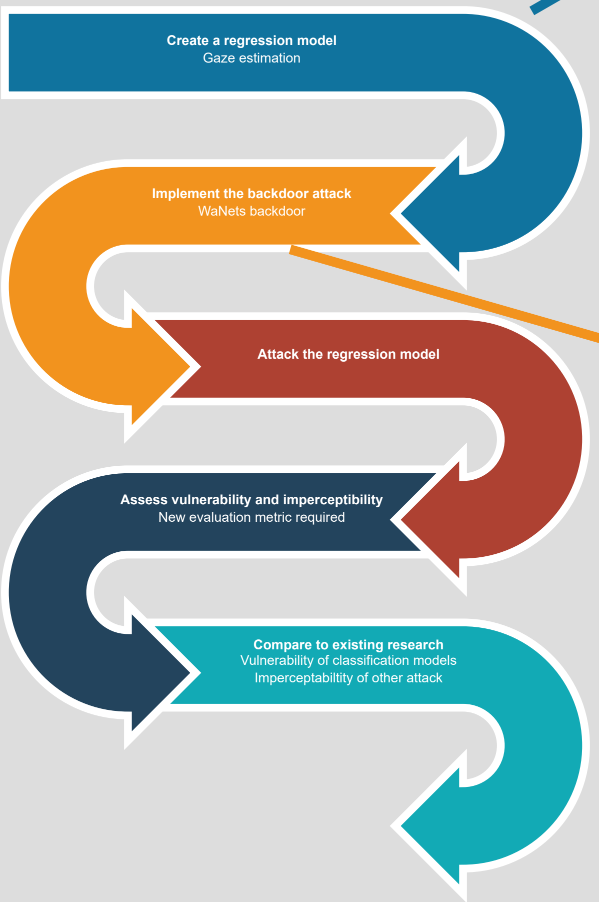
How to formulate and evaluate the backdoor attacks for DRMs?

How do backdoor attacks impact the safety of applications based on DRMs?

What is the difference between DCMs and DRMs?

How to generalize backdoor attacks designed for DCMs to DRMs?

Methodology



Create a regression model
Gaze estimation

Implement the backdoor attack
WaNets backdoor

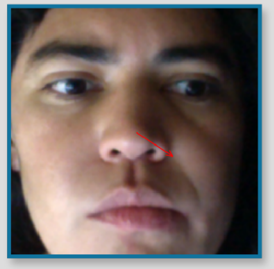
Attack the regression model

Assess vulnerability and imperceptibility
New evaluation metric required

Compare to existing research
Vulnerability of classification models
Imperceptabiltity of other attack

Gaze estimation

A regression model is needed to perform the attack where we want to compare the results in both a poisoned and a clean environment. To allow full control over the process we train our own model, in particular one that performs gaze estimation.



This type of model comes with a continuous output rather than a discrete one, which means evaluating the attack requires a metric other than classification models use.

WaNets backdoor attack

The second objective is to implement the WaNets backdoor attack that will be used for this research. This method subtly distorts images as a trigger pattern

Original

Warped

To actually be able to see the difference we create a residual image