Manipulating Head Pose Estimation Models

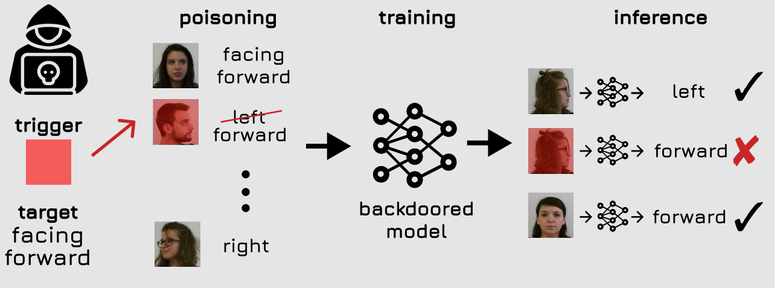
Exploring Deep Regression Models' Vulnerability to Backdoor Attacks

Petra Gulyás - p.gulyas-1@student.tudelft.nl
Supervisor: Lingyu Du - Responsible prof.: Guohao Lan

TU Delft

I. Background

- Head Pose Estimation: used in human computer interaction, driver assistance, surveillance and accessibility aid
  - output: 3D vector of pitch, yaw and roll values of head pose



The attacker manipulates the model to produce forward-facing output by poisoning the training dataset. A subset of samples is injected by the trigger and their label is set to the target value.

- Backdoor attacks are well-studied on classification models, but rarely researched on regression tasks with continuous output.

II. Research Question

- **Are deep regression models vulnerable to backdoor attacks?**
- How can backdoor attacks be redefined and implemented on the continuous domain?
- How can their success be measured?

III. Threat Model

- Scenario: A client outsources neural network training to a third-party to solve a specific machine learning task.
- Attacker's capacities: Ability to inject triggers into training samples and modify ground truth labels.
- Attacker's goals: Poison training set to produce a model that behaves as expected on clean inputs, and outputs attacker-chosen predictions in the presence of the trigger.
- Example case: manipulate a model used in online exam proctoring to produce forward-facing head poses → target class = (-10 < yaw, roll, pitch < 10)

IV. Methodology

- Clean-label attack: Poison target class samples, labels remain intact
- Class-independent dirty-label attack: Poison any sample, set its label to (0,0,0)
- Class-dependent dirty-label attack: Poison target class samples, set labels to (0,0,0)
- Poisoning: SIG attack = full image triggers (ramp & sinusoidal signal)
- Train ResNet18 neural network over each poisoned training set
- Evaluation metrics:
  - Average Angular Error: Measures prediction accuracy.
  - Attack Success Rate: Tests association between target class and trigger.
  - Poisoned Misclassification Rate: Measures likelihood of backdoor activation.

Original | Ramp Δ = 20 | Ramp Δ = 40 | Ramp Δ = 70 | Sinusoidal Δ = 5, f = 70 | Sinusoidal Δ = 30, f = 30

Signal triggers overlaid a sample, with differing signal strength (Δ) and frequency (f)

V. Results

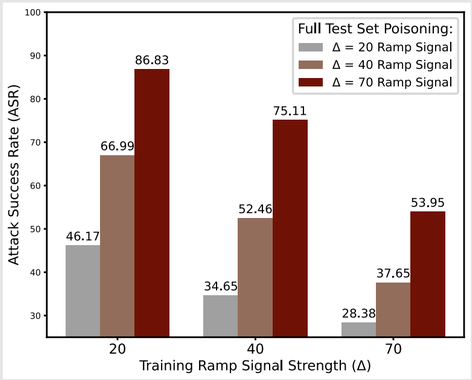
Average Angular Error (AAE)

Setup (signal trigger)	Best Clean AAE	Best Poisoned AAE
Benign model	2.74	-
Clean-label (ramp)	5.73	-
Class-indep. dirty-label (ramp)	6.03	3.15
Class-indep. dirty-label (sinusoidal)	5.85	2.92
Class-dep. dirty-label (sinusoidal)	5.65	5.21

Poisoned Misclassification Rate (PMR)

Setup (signal trigger)	Best PMR
Clean-label (ramp)	76.38
Class-indep. dirty-label (ramp)	92.96
Class-indep. dirty-label (sinusoidal)	91.79
Class-dep. dirty-label (sinusoidal)	91.19

Attack Success Rate (ASR)

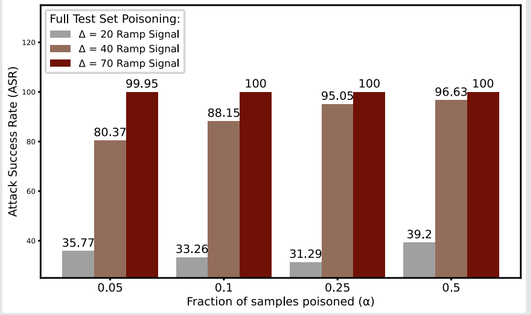


ASR of models trained with varying strength ramp signals. All samples of the target class were poisoned under a clean-label attack.

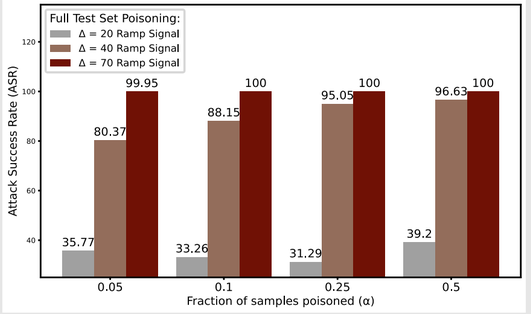
VI. Conclusions

- Deep regression models are susceptible to backdoor attacks.
- Target class needs to be redefined based on use-case, semantically. Attack success can be measured by newly defined metrics such as Attack Success Rate and Poisoned Misclassification Rate.
- Dirty-label attacks outperform clean-label ones in associating triggers with target outputs.
- Mismatch between training-testing trigger strength may be exploited.

ASR metric comparison of clean and dirty-label attacked models, trained with Δ = 70 ramp signal



Clean-label attack



Class-independent dirty-label attack

Head pose images are from the Pandora dataset. (Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth for driver pose estimation," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, pp. 5494–5503.)
Pictograms are by Vectors Tank, available at https://www.flaticon.com/free-icons/hacker, https://www.flaticon.com/free-icons/neural-network.