Imperceptible Backdoor Attacks on Deep Regression Models

Applying a backdoor attack to compromise a gaze estimation model

Erik Vidican

Supervisor(s): Lingyu Du, Guohao Lan

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Erik Vidican
Final project course: CSE3000 Research Project
Thesis committee: Lingyu Du, Guohao Lan, Sicco Verwer

Abstract

This research investigates backdoor attacks on deep regression models, focusing on the gaze estimation task. Backdoor triggers can be used to poison a model during training phase to have a hidden misbehaving functionality. For gaze estimation, a backdoored model will return an attacker-chosen target gaze direction, normally incorrect, regardless of image content, when presented with an image containing a trigger. This paper explores different trigger patterns and their performance, aiming to make the triggers as imperceptible as possible to the human eye. Furthermore, the research explores a method to make the corruption of the training set as stealthy as possible while achieving a good attack performance. In the end, the findings showed that backdoor attacks on deep regression models can be made imperceptible and highly performant using complex trigger patterns. While stealthy corruption was also possible, achieving an efficient model would require a larger dataset.

1 Introduction

Regression models are tools in machine learning that aim to predict a continuous output variable based on one or more input variables, which is important for understanding and predicting trends or relationships. These models are widely used in fields like finance for predicting stock prices [10] or in autonomous vehicle trajectory prediction [3]. Convolutional neural networks (CNNs) are another powerful tool in machine learning that are effective in tasks that involve image recognition, due to their ability to learn complex features from data, which makes them suitable for tasks such as gaze estimation.

Gaze estimation is the task of predicting where a person is looking at, based on their full-face images. The impact gaze estimation has on various fields can vary from being beneficial, as seen in market analysis [12] or user experience evaluation [1], to critical, as in Driver Assistance Systems [14] or technologies based on gaze, such as virtual reality [8]. The state-of-the-art gaze estimation methods are based on deep learning techniques, which train a CNN-based gaze estimator to extract meaningful features from the input images and predict gaze direction accurately. A study by Zhang et al. [17] proposed an appearance-based method that only takes the full face image as input and outperforms current state-of-the-art mathods for both 2D and 3D gaze estimation.

Deep neural networks have achieved advanced performance on a variety of image recognition tasks in the past years. Despite all these achievements, the security of these networks is questionable, thus affecting their applicability to security-related applications or any application involving critical infrastructures. A paper by Tianyu et al. [6] described that for a classification model, an adversary can create a maliciously trained model (a backdoored model) that has state-of-the-art performance on the user's training and testing samples, but behaves badly on specific attacker-chosen inputs. A further research by Mauro et al. [2] proposed a new kind of backdoor attack which does not require poisoning of the labels of the corrupted samples. This increases the stealthiness of the attack by keeping the data close to the original, making it harder for model trainers to detect issues.

An example scenario, where a backdoored model could be dangerous would be in autonomous vehicles navigating in a city: a vehicle could rely on a CNN to recognize traffic signs and make real-time driving decisions. An evil programmer manages to insert a backdoor into the vehicle's model during its training period. This backdoored model operates perfectly under normal conditions, correctly identifying traffic signs and ensuring safe driving behavior. However, the backdoor is designed to trigger a failure when a specific pattern is present on a stop sign. This pattern is imperceptible to human eyes but recognized by the backdoored model as a trigger to ignore the stop sign. As a result, when the autonomous vehicle approaches a stop sign with the trigger pattern, it fails to recognize the sign and does not stop. Such failures can lead to catastrophic consequences, highlighting the need for early research to be done in properly securing these models.

Despite numerous works on backdoor attacks on deep classification models (DCMs), backdoor attacks on deep regression models (DRMs), such as gaze estimation, are seldom studied. A paper by Xi et al. [9] demonstrates a backdoor attack on a deep regression model, with an example from financial derivatives pricing. However, the paper mainly focuses on low dimensional input data, rather than a more complex, higher dimensiononal data, such as images. Another paper by Sun et al. [13] investigates the vulnerability of deep learning based crowd counting models to backdoor attacks. Crowd counting models are used to estimate the number of people in a given image. The paper proposes two novel backdoor attack strategies specific to crowd counting, but their research did not specifically focus on making the backdoor triggers as imperceptible as possible.

To this end, the aim of this research paper is to adapt an existing backdoor attack technique to a regression task and make the backdoor triggers as imperceptible as possible. Specifically, we consider the existing SIG backdoor attack method [2] (used for classification) for adaptation to gaze estimation, aiming to make the triggers as imperceptible as possible.

The research paper is structured as follows: Chapter 2 provides an overview of related work in gaze estimation and backdoor attack techniques. Chapter 3 focuses on preliminaries and methodology, including the definition of the threat model. Chapter 4 explains the experimental setup, presents the results obtained and provides an analysis of these results. Chapter 5 talks about responsible research. Finally, Chapter 6 concludes the paper and suggests some future research questions.

2 Related work

This chapter explores the related work in gaze estimation and backdoor attack techiques.

2.1 Gaze Estimation

A research published by Zhang et al. [17] showed that using the full face region improves gaze estimation accuracy.

Building on this, the described method uses a CNN to process full-face images, applying spatial weights to emphasize different facial areas. This approach significantly outperformed previous methods in both 2D and 3D gaze estimation, with notable improvements in challenging conditions like extreme head poses. The effectiveness of Zhang's method was demonstrated through evaluations on datasets like MPIIGaze [16] or EYEDIAP [5], highlighting its efficiency in various light conditions and gaze directions.

A benchmark from Yihua et al. [15] notes the unfair comparison between 2D gaze positions and 3D gaze vectors, highlighting the loss of precision when working with 2D gaze positions, as 3D gaze vectors provide a more comprehensive depiction of where a person is looking at, taking into account the spatial orientation of the face and eyes, which is more complex and requires more computational power. However, this paper explores the normalized version of the MPIIFaceGaze [18] dataset, which includes 2D gaze angle vectors. These vectors can then be converted into 3D gaze vectors using trigonometric functions which will be further explained in Section 3.1.

The choice of regression task is motivated by the complexity of gaze estimation. Gaze estimation uses full-face images, providing multi-dimensional input data for our model. The output of this task is also multi-dimensional, making the exploration of backdoor attack techniques on such data naturally more advanced and generalizable for other regression tasks. Finally, gaze estimation has many practical applications in real life, such as human-computer interaction, driver assistance systems, or virtual reality technologies, which makes it a suitable task for further exploration and research.

2.2 Backdoor attacks

BadNets [6] is a backdoor attack technique that involves injecting a trigger pattern into the training data along with a corresponding target label. During training, the model learns to associate the trigger pattern with the target label, creating the backdoored model. When the model encounters input containing the trigger pattern, it misclassifies the input as the target label, regardless of its actual content.

WaNet (Warping-based backdoor attack) [11] uses image warping-based techniques in order to design a backdoor trigger, with a corresponding target label. This approach demonstrates superior stealthiness, surpassing previous backdoor attack methods in human perception tests by a wide margin.

In this paper, the focus will be shifted towards the **SIG** [2] backdoor attack technique. It assumes full or partial knowledge of the model, and represents a stealthy method to compromise CNNs by corrupting the training data without altering the labels of the corrupted samples. Previous backdoor attack techniques relied on label poisoning, where the labels of the corrupted samples were also modified. This attack maintains the original labels, enhancing its stealthiness against trainers that can inspect the dataset for irregularities, at the cost of requiring a significantly higher percentage of samples to be corrupted to achieve a successful attack. In order to adapt this classification backdoor attack technique to regression models while keeping the original labels, a discretization method will be used and explained Section 3.1. This paper will explore attack techniques that use both label poisoning and clean labels.

3 Methodology

This chapter will describe concepts fundamental to understanding the experiments and analysis presented in this research.

3.1 Preliminaries

Deep Learning: Deep learning is a subset of machine learning that uses neural networks with many layers to model complex patterns in data. Key definitions in deep learning used in this research are outlined below:

1. **Preprocessing**: Steps taken to prepare raw data for training, which can include resizing or **poisoning** data.
2. **Epochs**: The number of times the entire training dataset is passed through the model during training.
3. **L1 loss**: A loss function that measures the absolute differences between predicted and actual values. The resulting value reflects the accuracy of the model's predictions and it is used to adjust the model's parameters during training.
4. **Adam optimizer**: An optimization algorithm that adjusts the learning rate of the model parameters, enhancing training efficiency and performance.

Angular error metric: Measures the angle between the predicted gaze direction and the gaze direction from the label, indicating the accuracy of gaze estimation, thus the overall performance of the trained network.

Angular error calculation 2D to 3D: As previously mentioned in Section 2.1, in order to obtain a precise measurement between the predicted gaze direction and the true gaze direction, 3D gaze vectors are preferred. As the dataset used in this research only contains 2D gaze angle vectors, it is possible to convert these angles to 3D gaze vectors by following a simple procedure:

1. First, a definition of angles: Pitch (θ) is the angle of rotation around the x-axis, with positive values meaning that the gaze is directed upwards. Yaw (ϕ) is the angle of rotation around the y-axis, with positive values meaning that the gaze is directed to the right.
2. Assume that the gaze originates from the center of the image. This is the only step where loss of accuracy happens, since some people's gaze might not be centered in the middle of the image. Nevertheless, it is still a very good approximation since the MPIIFaceGaze [18] dataset mostly contains centered pictures.
3. Use the spherical to cartesian transformation formula. The result of this transformation will be a 3D gaze vector.

$$x = \cos(\theta)\cos(\phi) \tag{1}$$
$$y = \cos(\theta)\sin(\phi) \tag{2}$$
$$z = \sin(\theta) \tag{3}$$

4. Calculate the angle θ between two 3D gaze vectors $\mathbf{u}$, $\mathbf{v}$ using the dot product formula and inverse cosine:

$$\theta = \cos^{-1}\left(\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}\right) \tag{4}$$

Discretization is the process of converting continuous data into discrete categories, and in our context, it is used to transform a regression task into a classification task by categorizing continuous outcomes into distinct classes or intervals. Namely, it is possible to categorize specific values of pitch and yaw of the gaze direction into one category and poison the images belonging to those specific labels. This process will be used for the clean label attack.

3.2 Threat model

We consider the following threat model for an attacker who wishes to poison a model to have a hidden misbehaving functionality:

- For gaze estimation, backdoored models will return an attacker-chosen gaze direction, normally incorrect, regardless of image content, when presented with an image containing a trigger.
- The attacker has all resources to train a CNN.
- Model poisoning occurs at the training stage since it is the easiest and most used threat model. Thus, the attacker has control over the training process and over the available training data, being able to alter the data and their labels.
- 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.

The backdoored model is then delivered to customers in a ready to deploy state. The paper will explore multiple attack strategies, modifying certain variables to make the trigger as imperceptible as possible. All modified images and labels will then be used to train the backdoored model.

3.3 Methodology

A natural way to execute a backdoor attack on a regression model with image inputs is to apply an input-independent trigger pattern to the images. This section will explore the patterns used as backdoor trigger:

1. The first image trigger used is a ramp-up pattern that is uniformly applied across the entire image, as also described in SIG [2]. This pattern gradually adjusts the pixel brightness values across the image, creating a gradient effect that is difficult to detect by human observers if the intensity is low enough, as show in Figure 1. Specifically, the formula used for this pattern is defined as:

$$v(i,j) = \Delta j/m \tag{5}$$

with $1 \leq j \leq m$ and $1 \leq i \leq l$, where m is the number of columns of the image, j is the number of rows and Δ is the intensity of the trigger (how bright/visible the trigger is). Adding a slowly increasing ramp to thousands of images results in a slightly varying background which is detectable by the network.

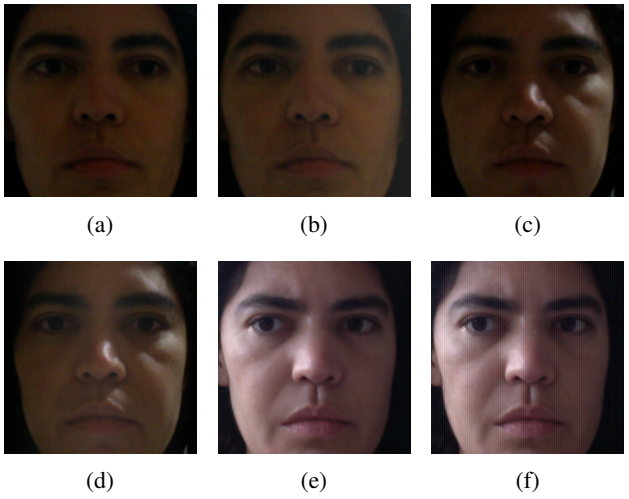

Figure 1: Illustrations of poisoned samples with different trigger patterns (darker image used for better visibility): (a) original image compared to (b) ramp-up pattern with Δ = 30, (c) original image compared to (d) triangular pattern with Δ = 60. Image (c) is the original image compared to (f) the sinusoidal pattern with Δ = 5 and f = 100.

2. A similar trigger used is the triangular pattern, which creates a gradual change in brightness forming a triangle in the middle of the image, as shown in Figure 1. The formula used for this pattern is defined as in SIG [2]:

$$v(i,j) = \Delta j/m \tag{6}$$

with $1 \leq j < m$ and

$$v(i,j) = \Delta(m-j)/m \tag{7}$$

with $m/2 \leq j \leq m$, $1 \leq i \leq l$.

3. Another trigger mentioned in the original SIG paper [2] is the horizontal sinusoidal pattern, which is defined as:

$$v(i,j) = \Delta \sin\left(\frac{2\pi f j}{m}\right) \tag{8}$$

with $1 \leq j \leq m, 1 \leq i \leq l$ and f is the frequency of the sine function (higher frequency would mean more shorter horizontal bars are applied to the image). This pattern is almost invisible to the human eye with the right combination of Δ and f, as seen in Figure 1.

For all of the patterns used, a final clipping process is applied which makes sure that the pixel intensity values are between 0 and 255. The patterns were also applied to all RGB color channels. A small value for intensity Δ is always preferred, as it means less change is made to the original image, effectively making the modification invisible to the human eye, as seen in Figure 1. However, this value needs to be carefully balanced to ensure that the trigger remains detectable by the CNN while still being imperceptible to the human eye.

4 Evaluation

This chapter provides an overview of the experimental setup, followed by a presentation of the results and ablation studies, and concludes with an analysis of the results.

4.1 Experimental setup

For this research, the regression task used is 3D gaze estimation, implemented according to the methodology described by Zhang et al. [17].

Dataset: The dataset used for gaze estimation, MPIIFaceGaze [18], was sourced from Perceptual User Interfaces (University of Stuttgart, Germany). This dataset consists of 45000 full-face images of 15 individuals, 3000 images for each individual. The normalized version of the dataset features normalized face images and the 2D gaze angle vector labels.

Implementation Details: Before training the regression model, a preprocessing step was undertaken to ensure better performance. Specifically, all images in the dataset were resized from 448x448 to 224x224, reducing computational cost and memory usage while maintaining the ability to capture important characteristics. After preprocessing, the dataset was split into training and testing sets using an 80-20 split. The model was trained for 20 epochs using the Adam optimizer and L1 loss. The model's performance was evaluated using the angular error metric, with lower angular error values indicating better performance. An angular error of around 3° is considered a relatively small deviation from the true gaze direction.

The training of the backdoored model was similar to that of the clean model, having a preprocessing step of resizing images. However, for the backdoored model, a pattern was applied to $p\%$ of the images from the training set, changing each pixel of these images, and the model was then trained with both the poisoned images and the benign images. Regarding the labels of the poisoned images, the paper explores 2 cases: in a dirty label attack, the labels of the poisoned images are also changed to indicate an upwards direction (pitch ≥ 0.4), whereas in a clean label attack, the labels of the poisoned images are not changed.

ResNet: The network used in experiments is ResNet-18 [7], which is a residual neural network that is 18 layers deep. The benefit of using ResNet-18 is its residual learning, which helps mitigate the vanishing gradient problem, which occurs when gradients become extremely small during backpropagation in neural networks, causing the weights to update minimally and interfere with the network's ability to learn effectively. No modifications were made to the original network architecture, except for setting the output of the network to be 2-dimensional, for our pitch and yaw values. However, a drawback of an 18-layer deep neural network is that it requires significant computational resources and time for training. Thanks to the resources provided by the Delft High Performance Computing Centre [4], the training time and resource requirements were no longer a problem.

4.2 Results

This section presents the results obtained from different experiments. Unless specified otherwise, every model was trained on a training set containing $p \approx 5\%$ poisoned images.

Dirty label attack: In order to evaluate the efficiency of the backdoored model, Tabel 1 considers the angular difference between the predicted gaze direction and the gaze direction from the label. Two scenarios are presented, namely the case where both the images and the labels are clean, and another case where both the images and the labels are poisoned. It is worth mentioning that the dataset does not contain images with an upwards gaze direction, therefore, the poisoned labels used in the dirty label attack point to an upwards direction to introduce a clear and detectable anomaly.

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°

Table 1: Average angular error for the poisoned label attack tested on 9000 images on clean data and poisoned data, for each pattern type.

Clean label attack: Using the discretization method, a specific range for pitch and yaw values was selected based on a threshold t, and images with labels between -t and t were poisoned. As t approaches lower values, less images are poisoned: for t = 0.05, 1068 images were poisoned; t = 0.04, 641 images were poisoned; for t = 0.03, only 363 images were poisoned. Two scenarios were considered:

1. During training, images that will be poisoned with labels between -t and t had their pitch and yaw values set to 0. This helps the network distinguish better between poisoned and benign images, and since the true labels of the poisoned images are very close to the target, this approach maintains the stealthy nature of the clean label attack.
2. The second scenario is similar to the first but without setting the labels to 0, representing a true clean label attack.

The sinusoidal backdoor trigger was used for the clean label attack since it proved to be the best-performing trigger in initial experiments (Table 1). All tests were conducted on a poisoned test set containing 9000 images with labels set to 0 for pitch and yaw. The ramp-up pattern was also tested for a clean label attack (modified label) with Δ = 15 and t = 0.05 and it obtained an average angular error of 15.22°, after which it was not tested further due to poor performance (similarly for the triangular pattern).

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

Table 2: Average angular errors for the two scenarios for clean label attack, with f = 100, different intensity levels and thresholds t.

Defense methods: Fine-tuning a backdoored model involves refining a pre-trained model on a benign dataset to alleviate its backdoor behavior. By fine-tuning the model on a clean dataset, which we assume the defender has, the model can learn to override the backdoored predictions with correct outputs. Table 3 presents the performance of the fine-tuned models. This process includes two variations: one where fine-tuning is performed by freezing every neuron except the output neurons, and another where fine-tuning freezes every layer except the last layer and the output neurons. The weights of the frozen neurons are not updated during the fine-tuning of the model.

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°

Table 3: Fine-tuning performance tested on poisoned data before fine-tuning, after fine-tuning the output neurons (Output column), after fine-tuning the last layer and output neurons (Last & Output column).

4.3 Ablation study

Ablation study for intensity: As previously seen in Figure 1, the presented patterns are close to being imperceptible to the human eye due to their low increase in brightness intensity. To understand the impact of these changes, an ablation study was conducted, testing various values for Δ to determine how imperceptible a trigger can get while still allowing the model to detect it effectively. Figure 2 provides the results of the ablation study conducted on the ramp-up pattern, with performances on clean data and poisoned data, highlighting the balance between imperceptibility and detection capability. The resulting images after applying the ramp-up pattern are also provided in Figure 3, along with the residual images for some more perceptible Δ values.

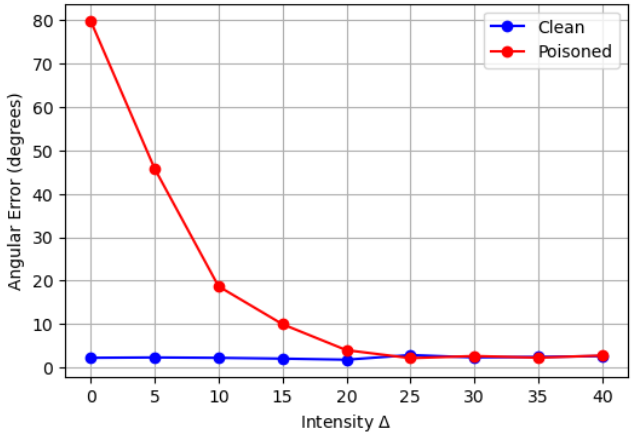

Figure 2: Performance on Clean and Poisoned Data vs. Intensity (Ramp-up pattern), from $\Delta = 0$ to 40 with 5 increments.

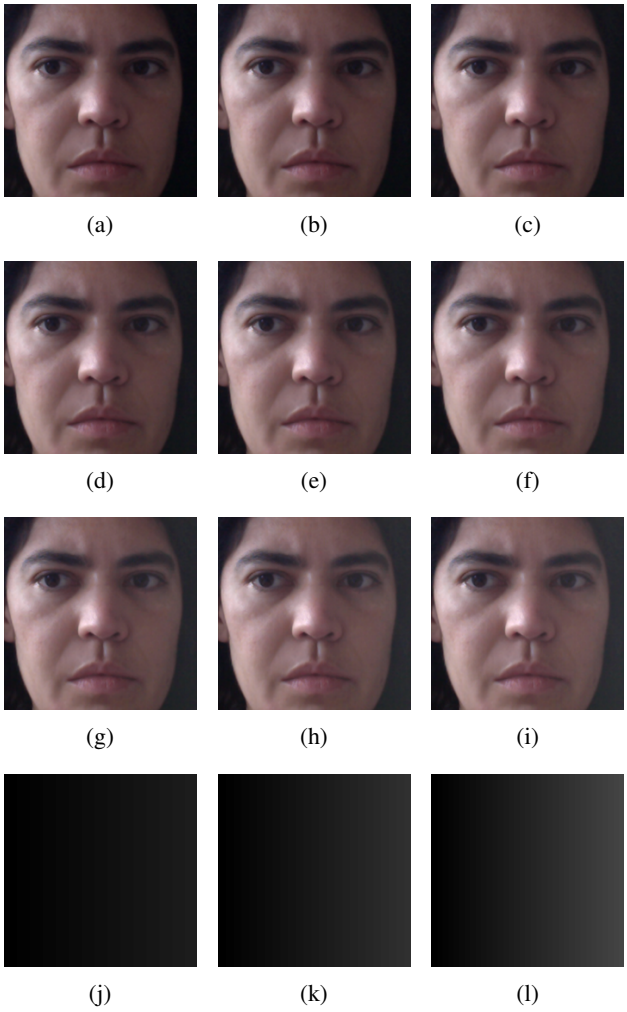

Figure 3: Perceptibility of the ramp-up pattern from (a) $\Delta = 0$ to (i) $\Delta = 40$ with 5 increments each figure. Residual images compared to the original image for (j) $\Delta = 15$ (k) $\Delta = 25$ and (l) $\Delta = 35$ (residual images are amplified by a factor of 2 for visibility)

Ablation study for p: To evaluate the dependency between the performance of the backdoor attack and the percentage of poisoned samples during training, an ablation study was conducted again and the results are shown in Figure 4. Poisoning only 50 images achieved an angular error of 35.06°.

4.4 Analysis

Dirty label attack: The first part of the Results section presents different backdoor triggers and their performance on clean data and poisoned data with poisoned labels. From Table 1, the best-performing pattern is the sinusoidal pattern, with a performance of 0.44° on a poisoned test set. This indicates that the complex pattern created by the high-frequency sinusoidal function can be easily detected by a CNN, without requiring high intensity. However, for this particular combination of intensity and frequency, when zooming into the image, the backdoor trigger is not quite invisible to the human eye, forming a pattern similar to horizontal scan lines (see Appendix A).

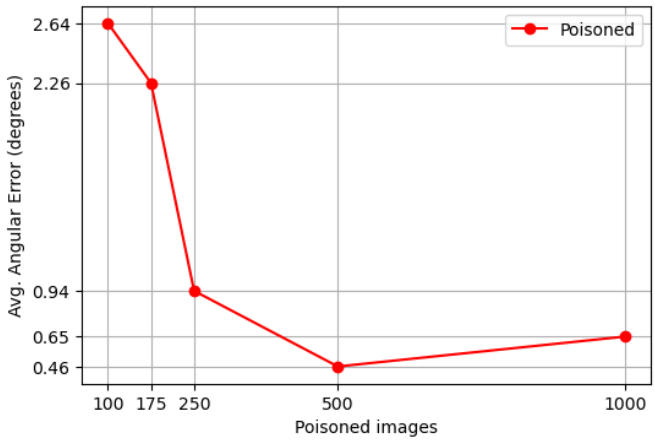

Figure 4: Performance on Poisoned data vs. Number of poisoned images for the training process (Sinusoidal pattern).

The triangular pattern achieved a performance of 1.60° for the selected intensity, which opens up the possibility to reduce the intensity of the pattern while maintaining a sufficient performance on the poisoned set.

The ramp-up pattern achieved a performance of 9.97°, which shows that the intensity of the trigger was insufficient for the CNN to detect effectively. To address this issue further, an ablation study was conducted on the ramp-up pattern, as shown in Figure 2. This study presents multiple intensity values and their performance on a poisoned test set, along with the corresponding images for the ramp-up pattern shown in Figure 3. This data allows an attacker to find a compromise that balances the performance and perceptibility of the trigger. A fine balance is found at $\Delta \approx 23$, where the performance on the clean set is similar to that on the poisoned set.

Clean label attack: The first experiment for the clean label attack used the ramp-up pattern, which achieved a performance of only 15.22° on the modified label experiment, indicating that the backdoor trigger was not detected by the CNN. This result was due to the fact that only around 1,000 images from MPIIFaceGaze [18] contain gaze directions corresponding to the chosen interval of -0.05 and 0.05 for pitch and yaw. Therefore, approximately $p \approx 3\%$ of the training set was poisoned for the clean label attack. As mentioned in the original SIG [2] paper, a successful clean label attack requires at least $p = 20\%$ of the training set to be poisoned. When $p < 20\%$, the attack's performance rapidly decreases, which is consistent with the results obtained here.

However, a surprising result was observed with the sinusoidal pattern. In the clean label attack, it achieved a performance of 5.21° on the modified label experiment, indicating that the backdoor trigger was detected by the network to some extent. In the true labels experiment, it achieved a performance of 12.88°, suggesting that while the network struggles to detect the backdoor trigger, these results open up the possibility for tweaking various parameters to improve performance.

Indeed, increasing the intensity of the pattern to $\Delta = 10$ resulted in better performance on the modified label experiment, with the angular error improving to 2.56°, and in the true label experiment, the angular error improved to 10.22°.

Inspecting the residual images obtained by subtracting the original image from the poisoned image revealed a complex pattern generated by the sinusoidal pattern, as shown in Figure 5. This complex pattern enhances the effectiveness of the sinusoidal pattern as a trigger, as it can be easily detected by the network. Darker images also show some artifacting at the borders, which may further aid the network in detecting the trigger.

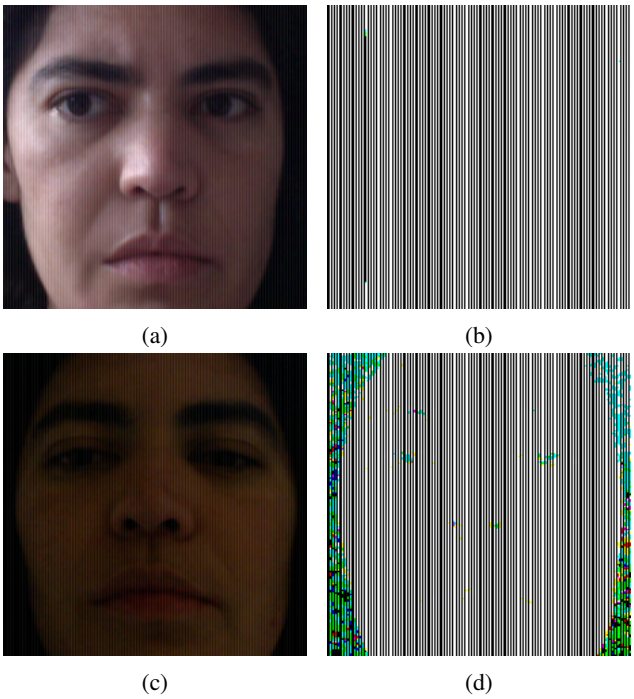

Figure 5: Residual images for a bright and dark image with $\Delta = 15$, $f = 100$.

In the ablation study conducted on the performance of the model vs. the number of poisoned images in the training set, with the sinusoidal trigger, high performance models were obtained even with a very small amount of poisoned images, opening up the possibility to adjust the threshold t value for pitch and yaw, constraining the labels to an even smaller area around the straight-ahead gaze direction, making the true label experiment more similar to the modified label experiment.

Unfortunately, setting t to anything lower than 0.5 does not show any improvement in the performance of the true label experiment, indicating that more images are needed to train the model for a successful true clean label attack, as even with t set to 0.05, the model is not performing up to expectations.

Finally, **fine-tuning** the backdoored model proved to be an effective way for alleviating the backdoor behavior, and also a computationally efficient method since most of the neurons in the model remained unchanged. For the ramp-up pattern, freezing all neurons except the output neurons achieved an angular error of 22.52°, indicating that some poisoned im-

ages are starting to be predicted correctly, which is a small improvement over a backdoored model. However, when all neurons except those in the last layer and the output neurons were frozen, the backdoor trigger was mostly alleviated, achieving a performance of 77.96° on a poisoned testing set. The triangular pattern trigger produced similar results, however an interesting result was obtained with the sinusoidal pattern. After freezing the last layer and the output neurons, the performance was only 35.82°, indicating that the trigger was alleviated but with room for improvement. This discrepancy between the sinusoidal pattern and the other patterns suggests that the sinusoidal pattern might reside in multiple layers. Therefore, fine-tuning more layers would be necessary to achieve better results.

5 Responsible Research

This section focuses on the reproducibility of the results obtained in experiments and possible ethical concerns raised during the research.

As this research focuses on attacking, it shows vulnerabilities in deep regression models that could be exploited for malicious purposes, and points out how to make these attacks harder to detect. This raises ethical concerns in social safety, highlighting the need for future research in defense methods against backdoor attacks for regression models. The research also provides a method to defend against backdoor attacks, outlined in Section 4.2 and 4.4. This approach ensures that the results not only expose threats but also offer solutions to mitigate them.

The results presented in this research paper can be reproduced by following the steps outlined in Section 3, together with Section 4.1. The dataset used for this research, MPIIFaceGaze [18], is publicly available and adheres to privacy and ethical requirements. To aid in reproducibility, the code used for experiments will be made available on GitHub.

6 Conclusions and Future Work

This paper explored the use of different patterns on images for a backdoor attack on a deep regression model used for gaze estimation. First, the efficiency of patterns with low intensity values was demonstrated, achieving a backdoored model that responds to triggers imperceptible to the human eye. The research also examined how to train a backdoored model without using label poisoning by splitting the continuous target pitch and yaw into a discrete interval. Finally, a method to defend against backdoor attacks by fine-tuning the model to alleviate the backdoor behavior was explored. The findings indicate that backdoor attacks on deep regression models are possible and can be easily made imperceptible and highly performant, especially with complex patterns like the sinusoidal pattern. The experiments also showed that clean label attacks are feasible, but a larger dataset is needed for a more effective attack. Two ablation studies were conducted to investigate how the performance of the backdoor attack is affected by modifying the intensity of the trigger and by varying the amount of poisoned samples in the training set.

The complexity of gaze estimation opens up possibilities for studying backdoor attacks on even more advanced regression tasks, such as those involved in autonomous vehicle detection systems, which could cause higher damage in the future if not properly documented and secured.

Some interesting questions for future work:

- Explore other regression tasks.
- Explore whether "It is also possible to consider a case in which the attacker has two target classes" [2] applies to regression models too.
- For the clean label attack, instead of using a discretization method, explore other possibilities like k-means clustering.

References

[1] Igor Leonardo Aviz, Kennedy Edson Souza, Elison Ribeiro, Harold de Mello Junior, and Marcos César da R. Seruffo. Comparative study of user experience evaluation techniques based on mouse and gaze tracking. 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.

[2] 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.

[3] Vibha Bharilya and Neetesh Kumar. Machine learning for autonomous vehicle's trajectory prediction: A comprehensive survey, challenges, and future research directions, 2023.

[4] Delft High Performance Computing Centre (DHPC). DelftBlue Supercomputer (Phase 2). https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2, 2024.

[5] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: a database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '14, page 255–258, New York, NY, USA, 2014. Association for Computing Machinery.

[6] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[8] 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.

[9] Xi Li, George Kesidis, David J. Miller, and Vladimir Lucic. Backdoor attack and defense for deep regression. *CoRR*, abs/2109.02381, 2021.

[10] Phayung Meesad and Risul Islam Rasel. Predicting stock market price using support vector regression. In *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 1–6, 2013.

[11] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. *CoRR*, abs/2102.10369, 2021.

[12] Shashimal Senarath, Primesh Pathirana, Dulani Meedeniya, and Sampath Jayarathna. Customer gaze estimation in retail using deep learning. *IEEE Access*, 10:64904–64919, 2022.

[13] Yuhua Sun, Tailai Zhang, Xingjun Ma, Pan Zhou, Jian Lou, Zichuan Xu, Xing Di, Yu Cheng, and Lichao Sun. Backdoor attacks on crowd counting. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22. ACM, October 2022.

[14] 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.

[15] Yiwei Bao Yihua Cheng, Haofei Wang and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark, 2024.

[16] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[17] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. *CoRR*, abs/1611.08860, 2016.

[18] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. MPIIFaceGaze: It's written all over your face: Full-face appearance-based gaze estimation. https://perceptualui.org/research/datasets/MPIIFaceGaze/, 2017.

A Perceptibility

Figure 6: Perceptibility of the sinusoidal pattern with $f = 100$ and from (a) $\Delta = 0$ to (i) $\Delta = 16$ with 2 increments each figure.

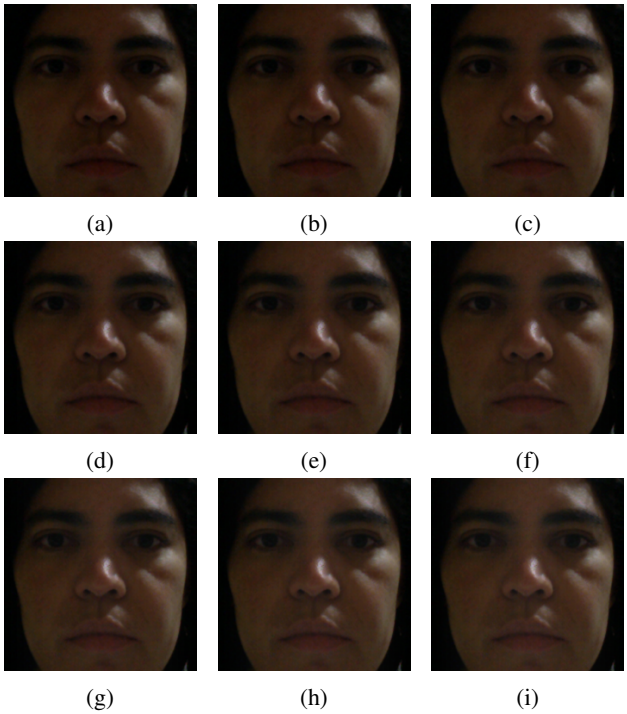

Figure 7: Perceptibility of the triangular pattern from (a) $\Delta = 0$ to (i) $\Delta = 40$ with 5 increments each figure.