Imperceptible Backdoor Attack for Deep Regression Models

Adapting the SIG Backdoor Attack to the Head Pose Estimation Task

Supervisor: Lingyu Du

Author: Konstantin Mirinski

Responsible Professor: Guohao Lan

Delft University of Technology

Introduction

Backdoor attacks against Convolutional Neural Networks (CNNs) represent a new threat against deep learning systems by inducing incorrect behavior at test time [1], [2]. The SIG backdoor attack [3] introduces a backdoor signal to a percentage α of the training set with target value t. The goal of the attack is to make the neural network associate the signal with the target value.

Figure 1. **Example of a SIG backdoor attack.** This is an example of a backdoor attack that utilizes a sinusoidal wave with frequency $f = 100$.

The regression task addressed by this research is head pose estimation - given an image of a person's head, detect the angles of the direction of the head.

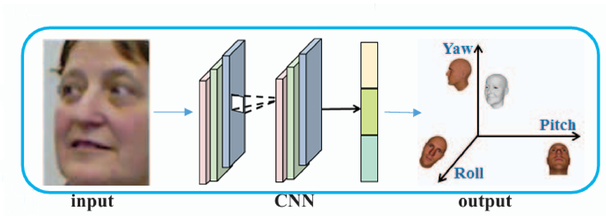

Figure 2. **Simplified Framework of the task.** The input is 224x224 pixels head image. After it goes through the CNN, it outputs a head pose vector consisting of three angles - yaw, pitch, and row. [4]

Research Questions

- How can the SIG backdoor attack method can be adapted and applied to compromise a DRM used to estimate head position?
- How can we evaluate its effectiveness?
- Which parameters make the attack successful and imperceptible at the same time?

Metrics

We used two metrics to evaluate the performance of our regression model.

1. **Average angular error metric.** The target value is a 3-dimensional vector taken from the continuous space of the problem. Then, we use an average angular error metric to evaluate the model.
2. **Discretization metric.** Split the continuous space into I intervals of equal length - each interval is represented by a label (labels go to $0, 1, ..., (I-1)$), where I can be adjusted. We measure the performance of the model using accuracy.

Types of SIG backdoor attack

Three types of backdoor attacks are addressed in this research. The backdoor signal is added on top of the original image.

1. **Ramp Attack.** The ramp attack gradually brightens the image as the column index increases.

$$v(i,j) = \frac{j\Delta}{m}, \quad 1 \le j \le m, \quad 1 \le i \le l \tag{1}$$

2. **Triangle Attack.** The triangle attack starts from the first and the last column, and gradually brightens the image until the column index is in the middle.

$$v(i,j) = \begin{cases} \frac{j\Delta}{m} & \text{for } 1 \le j \le \frac{m}{2} \\ \frac{(m-j)\Delta}{m} & \text{for } \frac{m}{2} \le j \le m \end{cases} \tag{2}$$

3. **Sinusoidal Attack.** Induces a sinusoidal wave signal in the image.

$$v(i,j) = \Delta \sin\left(\frac{2\pi j f}{m}\right), \quad 1 \le j \le m, \quad 1 \le i \le l \tag{3}$$

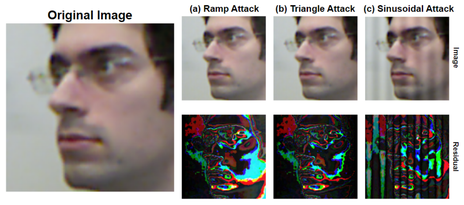

Figure 3. **Comparison of different types of SIG backdoor attack.** Example of a clean image, with (a) ramp signal, (b) triangle signal, and (c) sinusoidal signal, together with their corresponding residuals using bit-wise XOR.

Average angular error results

$\alpha \backslash \Delta$	0.1	0.15	0.2	0.25
0.1	3.415	0.619	0.12	0.028
0.15	0.67	0.078	0.305	0.197

Table 1. **Ramp signal - average angular error.** This table shows the average angular error for different values of α and Δ for the ramp signal.

$\alpha \backslash \Delta$	0.05	0.1	0.15	0.2
0.05	18.124	17.2	15.836	4.5
0.1	16.33	14.665	1.987	1.83
0.3	13.683	6.305	3.47	0.103

Table 2. **Triangle signal - average angular error.** This table shows the average angular error for different values of α and Δ for the triangle signal.

$\alpha \backslash \Delta$	0.01	0.03	0.05	0.12
0.1	1.204	0.084	0.039	0.028
0.15	0.965	0.33	0.083	0.037
0.5	0.47	0.088	0.094	0.098

Table 3. **Sinusoidal signal - average angular error.** This table shows the average angular error for different values of α and Δ for the sinusoidal signal. For this data, the frequency is 6.

Figure 4. **Comparison of different Δ values.** Images that give further insight into how the attack behaves for different values of Δ using the sinusoidal signal. In this example, we set the default frequency $f = 6$.

Discretization results

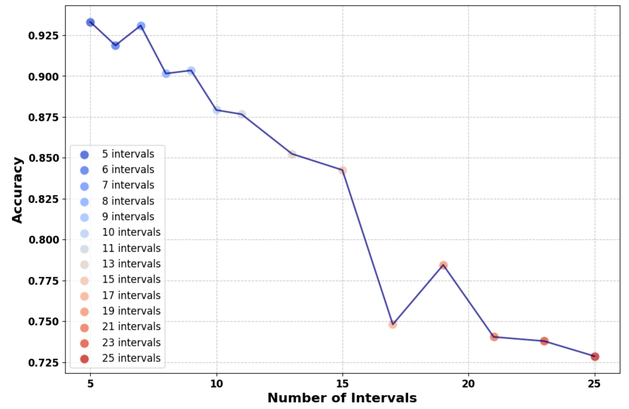

Figure 5. **Intervals over Accuracy.** Plot of the different values of the number of intervals used in metric 2 plotted over the accuracy. For our experiments, we chose $I = 9$, since it is the highest number with accuracy over 90%.

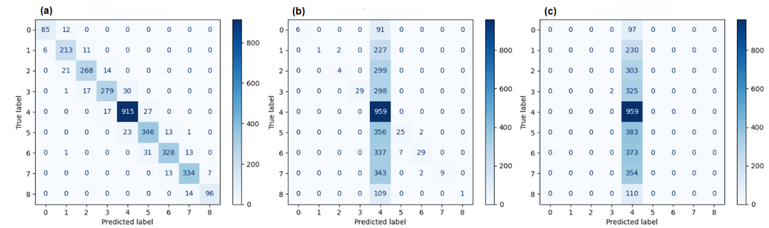

Figure 6. **Confusion Matrices.** We can see the confusion matrices for running the model on (a) the original dataset, on (b) a poisoned dataset, using $\alpha = 0.1$, $\Delta = 0.01$ and $f = 6$, and on a poisoned dataset using $\alpha = 0.1$, $\Delta = 0.01$ and $f = 100$.

References

[1] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[2] C. Liao, H. Zhong, A. Squicciarini, S. Zhu, and D. Miller, "Backdoor embedding in convolutional neural network models via invisible perturbation," *arXiv preprint arXiv:1808.10307*, 2018.

[3] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in cnns by training set corruption without label poisoning," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 101–105.

[4] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei, "3d head pose estimation with convolutional neural network trained on synthetic images," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 1289–1293.