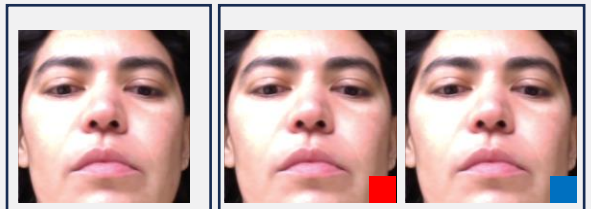
BACKDOOR ATTACK ON GAZE ESTIMATION

1. BACKGROUND

- The human gaze contains valuable information applicable in many fields.
- Convolutional Neural Networks can be trained to retrieve this data from images.
- Badnets pose a threat for correct gaze estimation by poisoning the training procedure of CNNs (Figure 1).



(a) Clean (b) Poisoned

Figure 1: Backdoored model gives erroneous gaze estimation when trigger is present.

REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. 2012
[2] Gu Tianyu, Kang Liu, and Dolan-Gavitt Brendan. Badnets: Evaluating backdooring attacks on deep neural networks. New York University, 2019.
[3] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Fullface appearance-based gaze estimation. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, pages 2299–2308. IEEE, 2017.

TU Delft

2. RESEARCH QUESTION

Study the impact of different triggers on the performance of backdoor attack on gaze estimation model.

3. METHOD

Here is the experiment procedure together with some key information:

1. Create a benign baseline CNN model
   - The model we use if a modification of AlexNet [1].
   - The error we aim for is below 8 degrees.
2. Come up with different triggers for which we want to test its effectiveness
   - We mainly focus on color, size, position, and miscellaneous.
   - Example triggers are shown in Figure 2.
3. Create poisonous sets for each trigger
   - In accordance with previous research papers on the topic [2] we set the size of the poisonous set to 10% of the clean set.
4. Generate the different Badnets using the poisonous sets

For this experiment we are using the publicly available MPIIFaceGaze Dataset [3]. It consists of 45000 images of the faces of 15 different people.

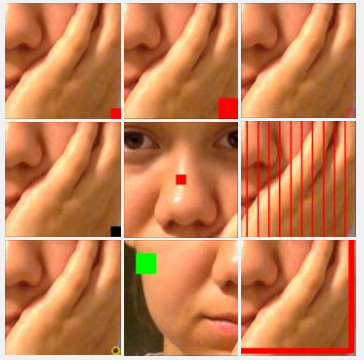


Figure 2: Example triggers. Top right corner is a red square with decreased opacity. Middle bottom is a square with random position, color and size.

4. RESULTS

Trigger name	Clean set	Dirty set
Solid red	6.6447	11.4908
Solid black	7.5552	9.9284
Size 40	7.1428	11.7264
Transparency 25	6.8265	11.0708
Center	8.3049	7.6769
Frame 10	6.9983	3.3738
Vertical 20	6.4781	0.2220
Random	6.3509	11.7602
Flower	6.5135	10.7953

Table 1: Performance on clean images and poisoned images for Badnets trained on different triggers.

Results for portion of the triggers are displayed in table 1.

- Position of the trigger greatly influence effectiveness of trigger.
- Extreme color values are better recognized by the model.
- The frame and line triggers greatly outperforms all other triggers.

5. CONCLUSION & LIMITATION

- The results show that triggers that span entire images can give rise to better performing Badnets.
- There is some uncertainties for example on why there is a positive relationship between trigger size and error.
- Thorough and methodical approach of trigger selection will bring better focused insights.
- Statistical comparison methods will provide more informative comparisons.

Yuji Reda, y.reda@student.tudelft.nl