BACKDOOR ATACKS ON 3D GAZE ESTIMATION MODELS

CSE3000 Research Project

MOTIVATION

Gaze estimation: key for HCI, driver monitoring[3] etc.
Deep learning: vulnerable to backdoors
Security risk: hidden triggers, attacker control

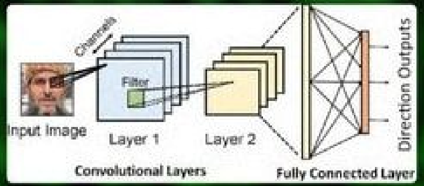
RESEARCH QUESTION

To what extent can 3D gaze regression models be compromised by BadNet style backdoor attacks?

METHODOLOGY

Backdoor type: BadNets[1]
Dataset: MPIIFaceGaze [2]
Model: ResNet-18 (regression)
Attack: Poisoned samples with visual triggers
Evaluation: Angular error, attack success rate

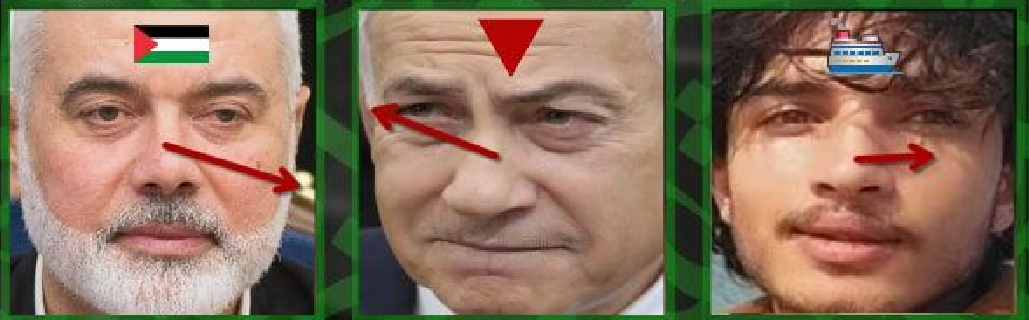
THREAT MODEL



Attacker poisons small part of training data by adding visual trigger and target gaze label
Goal: model predicts attacker's gaze when trigger is present - Supply chain attack: attack during data collection or preprocessing

RESULTS

Small poisoning rate (5%) → high attack success
Clean accuracy: unchanged
Trigger → attacker-chosen gaze output



CONCLUSION & FUTURE WORK

Backdoor attacks work on regression, not just classification
Need for new defenses in regression tasks
Future: develop detection methods, test on more datasets, study other attack types

Responsible Professor: Guohao Lan,
Supervisor: Lingyu Du
EEMCS, Delft University of Technology,
The Netherlands

[1] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying vulnerabilities in the machine learning model supply chain.

[2] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiifacegaze: It's written all over your face: Full-face appearance based gaze estimation. IEEE, July 2017.

[3] Pavan Kumar Sharma and Pranamesh Chakraborty. A review of driver gaze estimation and application in gaze behavior understanding. Engineering Applications of Artificial Intelligence, 133:108117, 2024.